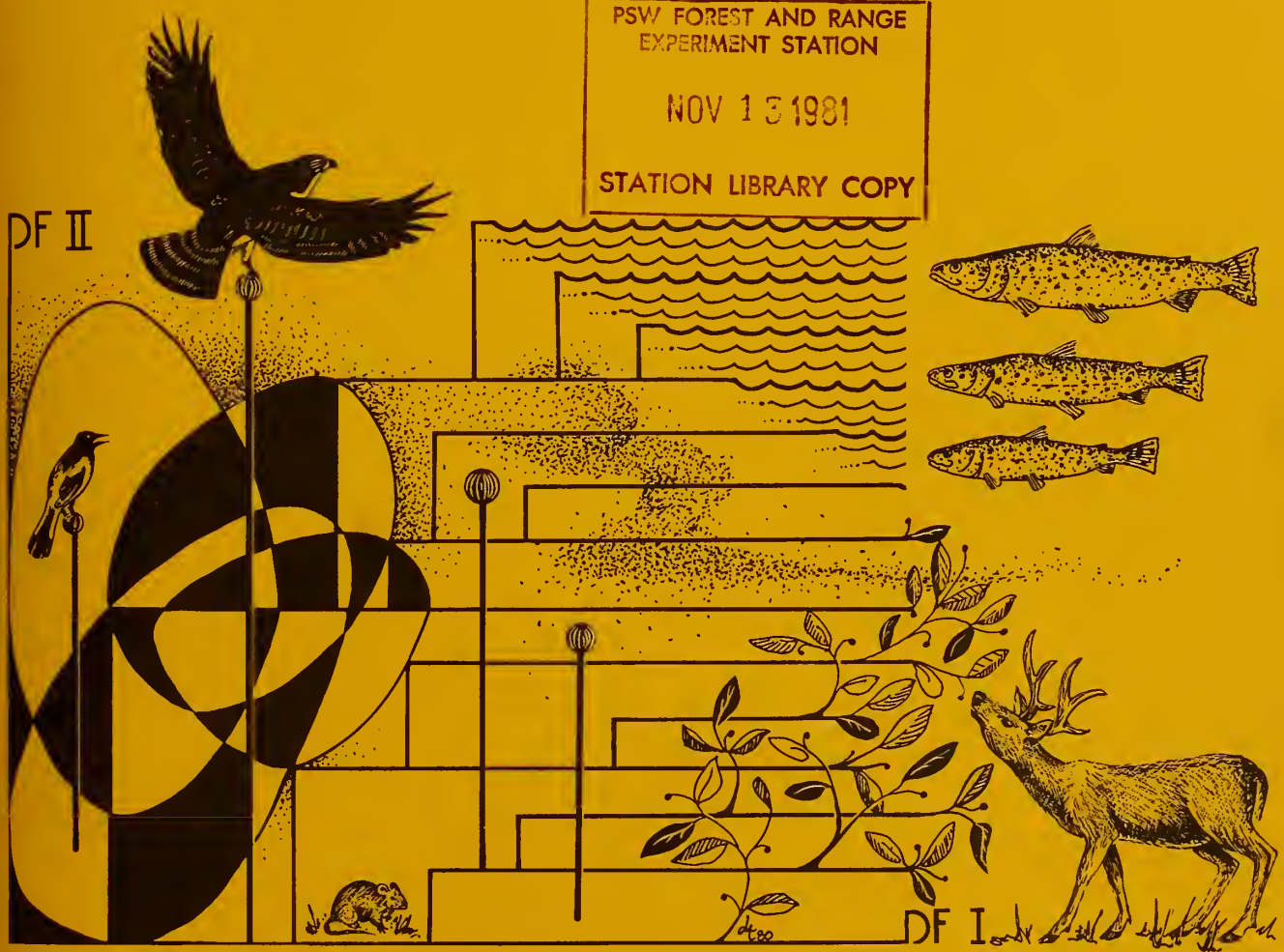
The Use of Multivariate Statistics in Studies of Wildlife Habitat



General Technical Report RM-87
Rocky Mountain Forest and
Range Experiment Station
Forest Service
U.S. Department of Agriculture

ABSTRACT

This report contains edited and reviewed versions of papers presented at a workshop held at the University of Vermont in April 1980. Topics include sampling avian habitats, multivariate methods, applications, examples, and new approaches to analysis and interpretation.

Limited numbers of reprints are available from the authors of each paper contained in this report.

USDA Forest Service
General Technical Report RM-87

August 1981

The Use of Multivariate Statistics in Studies of Wildlife Habitat[1]

Edited by

David E. Capen
University of Vermont

The workshop at which the papers included in this report were presented was jointly sponsored by

School of Natural Resources, University of Vermont

U.S. Fish and Wildlife Service

USDA Forest Service

[1]*This report was published by the Rocky Mountain Forest and Range Experiment Station. Supervision was provided by Thomas W. Hoekstra, project leader for the National Resource Analysis Techniques Project of the Resource Evaluation Techniques Program. Station headquarters is in Fort Collins, in cooperation with Colorado State University.*

ACKNOWLEDGMENTS

The editor acknowledges the following contributions to this report: Hugo H. John, Director, School of Natural Resources, University of Vermont, was an early supporter of the workshop on which this report is based, providing financial and staff support. Stanley Anderson of the U.S. Fish and Wildlife Service and Tom Hoekstra and Hans Schreuder of the USDA Forest Service served on the planning committee.

The following session chairmen evaluated submitted abstracts and reviewed manuscripts: James Karr, University of Illinois; Barry Noon, U.S. Fish and Wildlife Service; Robert Whitmore, West Virginia University; Jake Rice, Arizona State University and Memorial University of Newfoundland; Mark Boyce, University of Wyoming, and Ray Dueser, University of Virginia.

Graduate assistants and research technicians, Allen Boynton, Douglas Inkley, Stephen Parren, Bill Roberts, Douglas Runde, Mark Scott, Diane Tessaglia, and Frank Thompson contributed greatly to the organization of the workshop.

Trish McLaren, Lisa Klein, and Richard Lent provided editorial assistance. Richard Lindeborg rendered valuable service as the editorial liaison with the Forest Service.

Diane Tessaglia designed and contributed the cover artwork.

Maureen Douglas put all the manuscripts and revisions on the word processor and produced the camera copy.

PREFACE

The commonly recognized multidimensional nature of wildlife habitat has led to a rapidly increasing use of multivariate statistical techniques in studies of wildlife ecology. Techniques such as discriminant function analysis, principal component analysis, factor analysis, and canonical correlation have been applied in studies of habitat selection and resource partitioning, ordination of habitats and simulation of habitat change, and in development of habitat inventory systems. Although multivariate methods are not understood easily, they may be employed with little difficulty if one has access to a computer. Such convenience tempts researchers to employ sophisticated analytical techniques but overlook important statistical assumptions, experimental design, and biological interpretation. With these temptations in mind, a meeting was organized to bring research biologists and statisticians together to discuss multivariate methods and their applications to studies of wildlife habitat.

The meeting (held at the University of Vermont, Burlington, April 23-25, 1980) was called a workshop, although it was not unlike a research symposium. It was a working conference and participants took an active role in discussing and critiquing topics of concern. Biologists learned from statisticians and vice-versa. Interactive sessions and productive interchanges of ideas satisfied the workshop's purpose of encouraging the use of improved statistical methods in studies of wildlife habitat and fostering better interpretation of research results. It is hoped that these published proceedings will encourage other investigators to improve design of their research, analyses of their data, and interpretation of their results.

The papers presented at the workshop were either invited or submitted by abstract. Those papers are the basis of this report. Following the workshop, authors revised manuscripts and incorporated discussion from the meeting. Additional comments and critiques were recorded, edited, and added to papers where appropriate. Manuscripts were subsequently reviewed and revised so that these proceedings could be drawn together to form a cohesive volume, rather than a mere collection of papers.

This report was printed from camera-ready pages supplied by the University of Vermont, which is responsible for the accuracy and style of the contents. Statements of contributors may not necessarily reflect the policies of the U.S. Department of Agriculture.

CONTENTS

Examples: Multivariate Analyses of Wildlife Habitats

New Approaches to Analysis and Interpretation

SUMMARY OF THE WORKSHOP[1]

Stanley H. Anderson[2] and David E. Capen[3]

BACKGROUND

Wildlife habitat evaluation studies have progressed from purely descriptive work involving a discussion of community types and plant species present through concepts of vegetation function, physical structure, and vegetation structure. Many forms of special habitat descriptors, such as life forms or physiognomy, have been used.

Recently, resource agencies have been looking at means of evaluating wildlife habitat with ideas of describing changes that occur in wildlife populations as a result of habitat alteration. Five federal agencies (Forest Service, Fish and Wildlife Service, Soil Conservation Service, Geological Survey, and Bureau of Land Management) are working out a joint agreement for gathering and classifying wildlife habitat on a regional basis. Some federal agencies, many states, and some private organizations are developing data systems that include large amounts of wildlife habitat data. Objectives of these systems are to quickly classify wildlife habitat and indicate what changes might occur in wildlife as a result of widespread changes in the environment. Obviously, the utility of these systems depends on the type of data they contain.

Most papers presented at the workshop and included in these proceedings relate to species and groups of species. The application of wildlife habitat evaluation techniques has, therefore, jumped ahead of standardization of methods and development of analysis techniques. Development was the primary point of discussion in the workshop.

The computer age has revolutionized the analysis of research data. Statistical techniques are easily available to assist the biologist in examining relationships between wildlife and wildlife habitat. Techniques only theories two decades ago, are now employed routinely through the use of "canned" computer programs. Users of multivariate statistical methods requiring matrix algebra must rely on computers to analyze very large sets of data. Unfortunately, computer programs do not have control over either the quality of data analyzed or interpretation of results. Also, the researcher may have little knowledge of the proper type of program or method to be used for different forms of data. Yet multivariate analyses are becoming standard procedure in ecological studies.

CONTENT

Overview and Cautions

H.H. Shugart keynoted this meeting by pointing out that multivariate statistical techniques as methods of choice in analyzing habitat data among animals have three distinct advantages over alternative methodologies: 1) multivariate procedures intrinsically fit ecological problems (and data) of this sort; 2) many multivariate methods seem to be robust in the face of mild deviations from underlying assumptions; and 3) there already exists a hypergeometric interpretation of the relationship among animals (niche theory) that is essentially based on a multivariate sample space. He also discussed the need for more information on density, which has not been emphasized in most multivariate studies. His opening remarks initiated a recurring theme that biologists have not always planned statistical analysis prior to collecting data. Rather, they have sought statistical advice after massive amounts of data were gathered.

Douglas Johnson, an experienced biometrician, initially was invited to respond in a general sense to papers presented at the meeting. His response evolved into a comprehensive review of essential considerations which should <u>preceed</u> multivariate analyses; considerations such as non-linear response functions of wildife species to their habitats. This paper should be read before

---

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-35, 1980, Burlington, Vt.

[2]Leader, Wyoming Cooperative Fish and Wildlife Research Unit, University of Wyoming, Laramie, WY 82071.

[3]Assistant Professor, Wildlife Biology Program, University of Vermont, Burlington, VT 05401.

wildlife habitat studies are designed, and before multivariate methods are even selected to explore relationships or confirm hypotheses.

A more dramatic cautionary presentation was made by James Karr and Thomas Martin. These authors describe a study of bird/habitat relationships where habitat variables were reduced to "meaningful" axes by principal components analysis (PCA). Surprisingly, their habitat data were collected from a random numbers table, but findings compared favorably with those from published reports of real--we assume--bird/habitat studies. This paper emphasizes the importance of objective interpretation of PCA.

Habitat Measurements

A satellite session presented the opportuntiy to back away from data analysis and focus on habitats *per se*. This session, although restricted to avian habitats, provided both a theoretical and practical framework for designing studies of wildlife habitat relationships. The six speakers presented first a historical review of why we measure habitat and a theoretical perspective of how birds use multidimensional resources. Then there were discussions of how to select the proper habitat variables; how to measure them; and how to design statistical treatment for the data. Statisticians with little knowledge of biological processes will appreciate this chapter of the proceedings.

Theory and Methods

Biologists must have an understanding of multivariate statistical theory and methodology before employing these techniques in research endeavors. Papers given by Williams, Bhattacharrya, and Smith addressed the more commonly used methods employed in ecological studies. Ken Williams discussed problems in using discriminant function analysis (DA) in terms of habitat variables approching normality, problems of covariance equality in canonical analysis, and possible statistical violations found in the literature that have resulted in misinterpretation of data. His paper is of particular consequence in satisfying objectives of the workshop and the proceedings. Following Williams' presentation of the structure of canonical variates, Kim Smith gave an extensive review of the uses of canonical correlation in ecological work and formulated some important recommendations for improved use of this technique in our discipline. Research biologists are directed, in particular, to Smith's recommendations on sample size. Helen Bhattacharrya's presentation was an easily understood explanation of principal component analysis and its many variations. James Dunn prepared a particularly comprehensive treatment of transformations for univariate and multivariate data; this paper emphasizes the need for a thorough understanding of statistics before taking advantage of the more advanced options available in this area of quantitative science.

Application

Fourteen speakers gave papers which illustrated how multivariate techniques have been applied to field studies of a variety of different organisms. Andrew Carey uses PCA of a montane ecosystem to illustrate a suggested terminology for such ecological work. On a more applied basis, Tom Smith and co-authors Shugart and West show how important elements of a forest habitat can be identified by DA and incorporated into simulation models which predict habitat availability for selected species of birds. Chris Grue and co-investigators Reid and Silvy used stepwise multiple regression and DA to condense a large number of habitat variables into workable models for a large-scale classification and inventory system. Workshop participant Paul Geissler critiqued their study design by emphasizing the potential for prediction bias when the number of variables exceeds sample size. Like Karr and Martin, Geissler used a random number universe to illustrate his point.

Six of these 14 examples of multivariate applications may be studied as a collection of contrasting approaches to a variety of ecological problems. Martin Raphael used DA and cluster analysis to study nesting habitat of sympatric cavity-nesting birds. Mark Boyce and Joe Folse used canonical correlation in their studies of bird habitats. Boyce describes a robust analysis of sage grouse habitat, while Folse employs canonical correlation as an ordination technique, an unusual application. Brian Maurer and co-authors McArthur and Whitmore used PCA in what has become a common format in bird/habitat studies, but introduced a procedure of obtaining weighted mean values for habitat variables. Phil Sczerzenie applied PCA and PC-regression to deer harvest-land use relationships and illustrated the use of these techniques on an expanded spatial scale. Tom Harshbarger and Helen Bhattacharrya contributed the only paper where multivariate techniques were used in a study of an aquatic species, trout. Their conclusion was that regression models based on derived factors were no better than models composed of original variables.

New Approaches

Five research papers were particularly relevant to the purposes of the workshop; these presentations introduced both new statistical applications to familiar problems and new field approaches to familiar statistics. Jake Rice, R.D. Ohmart and B.W. Anderson illustrated the importance of seasonal and annual variation when using DA to classify avian habitats. The application of their findings cannot be overlooked. John Rotenberry and John Wiens also used a familiar approach, PCA, but reported an innovative technique of combining species abundance and habitat relationships in the same environmental space.

The remaining papers dealt with the important

topic of robust procedures. Jim Harner and R.C. Whitmore described new techniques which are robust to outlying data; their computer programs will be sought by many investigators. The problem of multicollinearity is addressed by Janet Cavallaro, J.W. Menke, and W.A. Willimas; their use of ridge regression to deal with this problem is applaudable. Multicollinearity is also highlighted in Kathryn Converse and B.J. Morzuch's paper.

Lyman McDonald, like Doug Johnson, is a statistician who has worked extensively with wildlife research problems. He too was asked to respond in general to papers presented at the workshop. Dr. McDonald reviewed papers and chose to concentrate on robust procedures, indicating the importance of this topic to wildlife habitat studies. His comments are found after those papers which address robust techniques.

CONCLUSION

Although insights often occur when techniques do not work, we must adhere to basic scientific premises. Multivariate analyses are useful tools for describing wildlife habitat only when properly applied; when misused or abused they become not only ineffective, but also disastrously misleading. It is important to remember that multivariate statistics alone do not solve problems; they only assist us in using our knowledge to interpret large amounts of data. We should not try to make sense out of nonsense; however, exploratory studies are valid when they follow the stated purpose.

To competently continue our work in examining wildlife habitat, it is most important for us to clearly define the problem at the outset of each study. This means setting objectives, establishing hypotheses that can be tested, and determining the forms of data to be collected and how they should be analyzed statistically. These steps will help determine data collection procedures. At this point it is extremely important to know the assumptions of tests to be used and make sure that these assumptions are met in the data collection procedure. Finally, we must recognize that results of the tests are not an end in themselves, they only provide direction. We must return to the field and validate the results, frequently through manipulative experiments. We as investigators, teachers, and scientific reviewers, must encourage proper interpretation through our selection of tests.

Site-specific studies must be used as one example of regional and national application. Researchers must be keenly aware of their need to design experiments and translate results into meaningful information for teachers, managers, and lawmakers. A cautionary question: are we publishing our results too soon? Maybe studies should go beyond the typical one or two years and proper verification should follow.

We hope this symposium of biologists and mathematicians is of help to all involved. The demands for the information we gather are great but it is important that we pass on only the best information. The type of interaction brought about by this workshop needs to continue.

AN OVERVIEW OF MULTIVARIATE METHODS AND THEIR APPLICATION TO STUDIES OF WILDLIFE HABITAT[1]

H.H. Shugart, Jr.[2]

Abstract.--Multivariate statistical techniques as methods of choice in analyzing habitat relations among animals have three distinct advantages over competitive methodologies: 1) Multivariate procedures intrinsically fit ecological problems (and data) dealing with habitat selection. 2) Many of the multivariate methods seem to be robust in the face of mild deviations from the underlying assumptions. 3) There already exists a hypergeometric interpretation of relations among animals (niche theory) that is essentially based on a multivariate sample space. These considerations, joined with a reduction in the cost of computer time, the increased availability of multivariate statistical "packages," and an increased willingness on the part of ecologists to use mathematics and statistics as tools, have created an exponentially increasing interest in multivariate statistical methods over the past decade. The earliest multivariate statistical analyses in ecology did more than introduce a set of appropriate and needed methodologies to ecology. These studies emphasized different spatial and organizational scales from those typically emphasized in habitat studies. The traditional wildlife habitat study was based on measuring the density of a population in a homogeneous plant community. New studies, using multivariate methods, emphasized individual organisms' responses in a heterogeneous environment. This philosophical (and to some degree, methodological) emphasis on heterogeneity has led to a potential to predict the consequences of disturbances and management on wildlife habitat. One recent development in this regard has been the coupling of forest succession simulators with multivariate analysis of habitat to predict habitat availability under different timber management procedures.

Key words: Habitat selection; multivariate statistics, quantitative ecology; succession models; wildlife-vegetation interactions.

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Senior Research Staff Member, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830.

INTRODUCTION:

THE EVOLUTION OF MULTIVARIATE HABITAT ANALYSIS

Investigators in research centers all over the world, and particularly in the United States, are discovering applications of multivariate statistical techniques in studying the habitat relations of a diverse range of animals and plants. Given the suitability of multivariate statistics for habitat analysis coupled with the need for a wider understanding of animal/environment relations mandated by man's increasd use of the earth's resources, this is a logical situation. Indeed, the logic of using multivariate analysis to manage animal habitat would seem to make such applications inevitable. Yet only a decade ago, no studies that are methodologically and philosophically equivalent to those of today were in evidence. This paper takes as its central tenet that multivariate statistical analysis of habitat requirements of animals is the product of a synthesis, occurring in the early 1970's at several different research centers, that united different lines of scientific research.

Figure 1 illustrates the main elements of this synthesis. Two developments that were independent of ecological studies, the increased availability of computer time on high-speed digital computers and the development of multivariate statistical techniques, were combined with three ecological developments: 1) the hyperspace theory of the niche, 2) the realization that small spatial-scale studies could reveal much

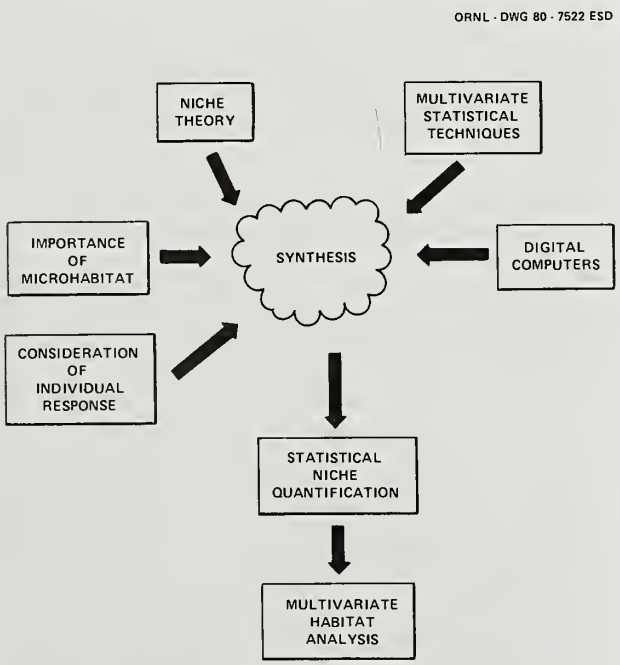


Figure 1. Schematic diagram of the scientific research elements that combined in a synthesis to produce multivariate habitat analysis.

on animal interrelations, and 3) an emphasis on individual organism response as being important in determining species distributions. A number of papers developed various aspects of what was to become multivariate analysis published before the 1970's, but most of these papers did not combine all elements of the synthesis (table 1). MacArthur's (1958) classic work emphasized a relation between individual microhabitat utilization and the hyperspace niche theory model (Hutchinson 1944, 1957, 1965). Klopfer (1965) developed an impressive line of research in associating individual organism responses to microhabitat, as did Wiens (1969). Cody (1968) and Fujii (1969) both developed the idea of using multivariate statistics in studies of animal niches. In 1971 and 1972 a number of papers, by individuals working independently, appeared that combined these elements and attempted to quantify niches of various organisms (Green 1971, Hespenheide 1971, James 1971, Martinka 1972, Shugart and Patten 1972). The management implications of these works became obvious to various individuals and within a few years several papers had appeared that discussed the use of multivariate habitat analysis as a management tool. Because of its synthetic origins, multivariate habitat analysis can be considered in terms of the elements that formed this synthesis and in terms of the problems intrinsic to these synthesis elements.

ELEMENTS OF SYNTHESIS

Quantitative Elements: Statistics and Computers

Two of the most important elements that led to the development of multivariate habitat analysis were the state of development of multivariate statistical procedures and the increased availability of high-speed digital computers. It is interesting to note that current user-oriented statistical analysis procedures (e.g., Cooley and Lohnes 1971, Dixon 1974, Barr et al. 1976) are further accelerating the "computerization" of ecological studies in general. Many of the statistical procedures used in multivariate habitat selection studies require rather involved numerical analysis programs that: 1) use computers and 2) would be difficult for ecologists to develop de nova.

The availability of multivariate statistical analysis programs and computer time created a situation in which the initial work in developing habitat analysis techniques in the 1970's was often done in conjunction with a consulting statistician (e.g., Dr. J.E. Dunn, a contributor to these proceedings). This has had a positive effect in that the rigor in testing to meet statistical assumptions, in using the "right" methodologies, and in interpreting results correctly is much higher in multivariate habitat analysis than in botanical ordination and classification procedures. Ecologists interested in plant ordination developed a number of analytical techniques with little rigor relative

Table 1. Papers involved with animal/habitat relationships before and during the synthesis period for multivariate habitat analysis. A "yes" to Niche theory indictes the paper is strongly oriented toward niche theory; "yes" to Microhabitat indicates a use of a small spatial scale sample size; "yes" to Individuals indicates an emphasis on individual organisms; "yes" to Management indicates management potential is discussed in paper.

Multivariate paper	Taxa	Methods	Niche theory	Microhabitat	Individuals	Management
MacArthur (1958)	Birds	-	Yes	Yes	Yes	-
Klopfer (1965)	Birds	-	-	Yes	Yes	-
Wiens (1969)	Birds	-	-	Yes	Yes	-
Cody (1968)	Birds	Discriminant function (DF)	Yes	-	Yes	-
Fujii (1969)	Insects	Principal components analysis (PCA)	Yes	-	Yes	-
Green (1971)	Molluscs	DF	Yes	Yes	-	-
Hespenheide (1971)	Birds	DF	Yes	Yes	-	-
James (1971)	Birds	DF, PCA	Yes	Yes	Yes	-
Martinka (1972)	Birds	DF	Yes	-	-	Yes
Shugart and Patten (1972)	Birds	DF, other techniques	Yes	Yes	Yes	-
Anderson and Shugart (1974)	Birds	DF, PCA	Yes	Yes	Yes	Yes
Green (1974)	Benthic Animals	DF	Yes	Yes	Yes	-
Shugart et al. (1974)	Birds	DF, PCA	Yes	Yes	Yes	Yes
Shugart et al. (1975)	Birds Mammals	DF	Yes	Yes	Yes	Yes
Conner and Adkisson	Birds	DF	Yes	Yes	Yes	

to underlying statistical assumptions. The difference in rigor between plant ordination and animal habitat analysis is quite pronounced to anyone who has worked in both areas. While this would appear to be a positive attribute relative to habitat studies, this is not necessarily the case as I will discuss below.

Niche Theory

Hutchinson (1944, 1957, 1965) formulated a hypergeometric model of the niche of an organism that captured the imagination of theoretical ecologists for several decades (e.g., Horn 1966, Maguire 1967, McNaughton and Wolf 1970, Colwell and Futuyma 1971, Pielou 1972, May 1974, 1975), although this geometric interpretation has tended to be lost from more recent interpretations. The hypergeometric concept of the niche and the n-dimensional sample space are analogous in many respects, and this analogy was noticed by several early investigators (Table 1). The existence of a set of theories (niche theory) has led to an extended level of interpretation of multivariate habitat studies. Thus, one can interpret lack of overlap in terms of theories of "limiting similarity" or "competitive exclusion" (Harner and Whitmore 1977, Dueser and Shugart 1978). Further, the pattern of the means in a sample or discriminant space can be interpreted relative to theories of community structure (Shugart and Patten 1972, Dueser and Shugart 1979).

Consideration of Individual Responses and the Importance of Microhabitat

Population-level censuses of bird populations have for years been coupled with plant community surveys to obtain a measure of species response to vegetation. While these surveys will continue to be important in the future as they have in the past, they do not necessarily provide much insight into the detailed aspects of why one location seems more suitable for a species than some other location. Interest in this microscale problem as well as the theoretical underpinnings of habitat selection had been published quite early (e.g., von Uexkull 1909, Kohler 1947, Tinbergen 1951, Harris 1952). In the mid-1960's, there was an intensified interest in habitat selection as a behavioral phenomena (e.g., Wecker 1963, Klopfer 1965, MacArthur and Pianka 1966). This emphasis on individual organisms tended to produce observational data sets with very high degrees of freedom and with more than one variable recorded for each observation. Such data created a need to

explore multivariate statistics as an analytical tool.

PROBLEMS INTRINSIC TO THE SYNTHESIS ELEMENTS

The synthesis of multivariate habitat selection methodologies proceeded from the combining of several elements of research from different fields in the 1970's. As a product of a rapid synthesis, today's procedures suffer from certain problems intrinsic to the component elements.

Statistical Assumptions Versus Niche Theory

Multivariate habitat selection studies have a strong emphasis on using the "correct" statistical procedures. Most letters and phone calls that I receive regarding habitat selection involve deciding what method to use, not on how to interpret analytical results. Further, as contrasted to plant ordination studies in which authors typcially reference their methods to other ordination studies, multivariate habitat studies refer their methods to reputable multivariate statistical textbooks. This healthy respect for statistics is probably quite good, but it also makes papers in the field proceed with ponderous inevitability. I am of the opinion that rigorous adherence to methodological correctness probably has prevented incorporation of the heuristically rich geometric theories of the hyperspace niche concept from being fully developed in wildlife habitat studies.

Inclusion of Population-Level Aspects

Most multivariate habitat analyses are so focused on the responses of individual organisms to microhabitats that concepts of animal density are almost lost. Some species appear to be quite uncommon even in the face of an apparent abundance of suitable microhabitat. The Swainson's warbler (Limnothlypis swainsonii) is a possible example of such a species. It is difficult to include population dynamics in a detailed microhabitat study, yet such dynamics are essential to management of habitat for selected species. I believe that this will be one of the challenges to workers in this field over the coming decade.

Inclusion of Macrohabitat Considerations

The consideration of the pattern of microhabitats at larger spatial scales (macrohabitat) is difficult to include in the present multivariate habitat analysis methodologies. In fact, macrohabitat considerations are often viewed as an unfortunate sampling situation in which the sample space has a pattern of internal clusters. There may be considerable importance in arrangement of elements of the landscape that are suitable for maintaining viable populations of a species. Rosenzweig's (1973, 1974), Rosenzweig and Winakur's (1969), and Schroder and Rosenzweig's

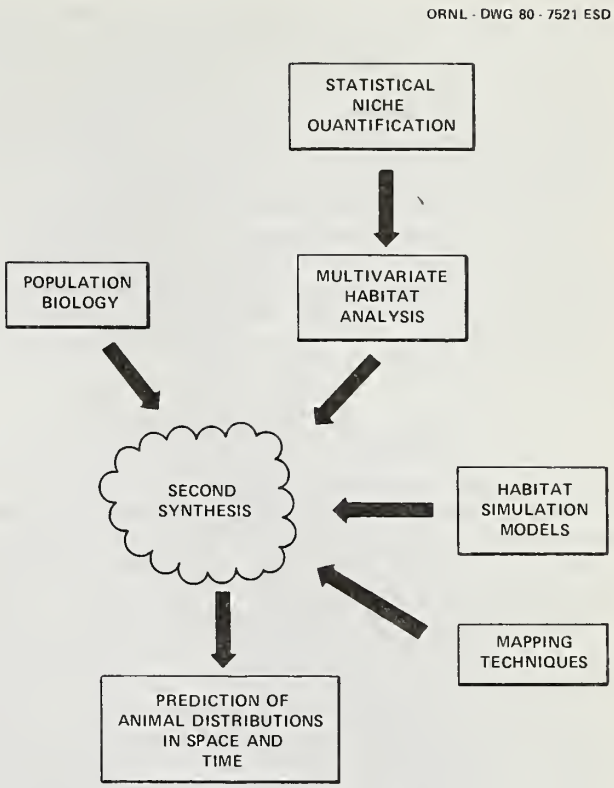


Figure 2. Schematic diagram of scientific research elements that may combine to produce a second synthesis extending from multivariate habitat analysis.

(1975) pioneering work on the theoretical and experimental responses of populations to altered patterns of microhabitats is an important benchmark set of studies that needs to be repeated in other communities.

THE ELEMENTS OF A NEW SYNTHESIS

In the previous section, I have tried to identify briefly some areas of new fruitful exploration for studies of habitat selection. There is a need to use the powerful methodology of multivariate habitat analysis to develop new theory, as well as to contine its use in difficult applications. In the former area, I believe that work I have been involved with in trying to develop a concept of niche patterns for communities (Shugart and Patten 1972) and particularly the development of this idea by my colleague Dr. R. D. Dueser (Dueser and Shugart 1979) are examples of attempts to extend from simple data analysis to theory. These proceedings hold great promise for stimulating other theoretical work and should definitely provide some classic examples of applications.

There is, in my opinion, a need for a second synthesis in the field and I would like to briefly

identify what may be the important elements of this synthesis (fig. 2). First, there is a need to meld the exciting developments made in population biology over the past decade with the equally exciting developments in multivariate habitat analysis. The potential for cross-seeding these lines of research is great, and the only limitation in uniting the field is in the formidable mathematical development that must be unified between the two. Two important quantitative developments (the ability to simulate changes in microhabitats through time and computer mapping of habitats) are already beginning to be included in habitat studies, and two examples are provided below.

Habitat Simulation Models

One interesting development in the field of ecosystem modeling over the past few years has been the development of forest succession simulation models capable of providing extremely detailed predictions on the future states of forests (reviewed by Shugart and West 1980). The level of detail and spatial scale of the output of some of these models is similar to that used in multivariate habitat selection studies. This convenient parallel development opens the possibility of projecting the temporal pattern of habitat availability following either man-made or natural disturbances. Figure 3 is an example of such an application using the Appalachian deciduous forest succession (FORET) model (Shugart and West 1977) to project habitat conditions for the ovenbird (Seiurus aurocapillus) on Walker Branch watershed, a site of intensive ecological investigation located on the Department of Energy (DOE) reservation in Oak Ridge, Tennessee. The FORET model simulates the birth, death, and growth of each tree on a 0.085-ha circular plot. By collecting habitat data on the presence of the ovenbird that has corresponding habitat information to that predicted by the model, one can use the model as a habitat simulator. In this particular example, we used discriminant function analysis to classify habitat versus non-habitat. As can be seen from an example, one can also use the detailed succession simulator to perform model experiments such as projecting the habitat dynamics of a 22.9 cm (9-in) diameter-limit cut of commercially valuable species (fig. 3). Details of this particular methodology will be treated later in this volume (Smith et al. 1981). I mention this example here to identify a need for adding dynamics to our currently largely static methodologies.

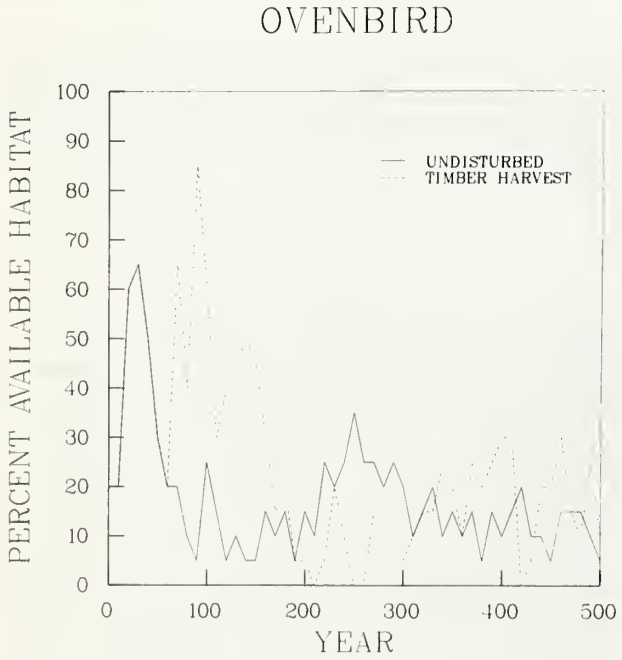


Figure 3. Ovenbird habitat availability over 500 years on Walker Branch Watershed under two treatments. Results are from the FORHAB model (Smith et al. 1981).

Mapping Techniques and Data Sets

There are presently a large number of data sets (e.g., the USDA Forest Service Continuous Forest Inventory [C.F.I.]) that could be used in conjunction with multivariate habitat data to make state- or continental-scale maps of distributions of wildlife habitat. The problems here involve keying habitat variables associated with a given species in a multivariate habitat study to the variables that can be obtained from inventory data sets. A fine example of this appraoch is Lennartz and McClure's (1979) application of C.F.I. data to map the potential extent of the red-cockaded woodpecker (Dendrocopus borealis) in the southeastern U.S. The determination of the level of variation over the continent of various species' habitat selection would be a valuable set of information for developing these maps.

THE SECOND SYNTHESIS

In this paper I have taken a broad view of the ontogeny of a still-developing understanding of habitat selection of several animals. The excitement of being involved in a young field is contagious, and I hope that this meeting will, by virtue of the increased interaction of scientists involved in multivariate habitat studies, help create a period of reviewed synthesis. The primary elements of such a synthesis might include the elements that I have mentioned (fig. 2), or they may take some entirely different direction -- only time will tell. Whatever the case, the present symposium should provide much fuel for the fire.

The healthiest aspects of multivariate habitat analysis have been of a philosophical rather than of a methodological nature. Studies have attempted to be rigorous in the sense of statistics while, at the same time, they have attempted to be both theoretical and explanatory.

This balance should not be lost. The studies have traditionally emphasized an understanding of ecological mechanisms at a fine scale of resolution. If such detail can be related to regional maps, the contribution to biogeography could be considerable.

In my opinion, the current research direction and velocity could, within the decade, provide such research products as dynamic maps of regional habitat availability for a great number of species, with the potential to determine the changes in these maps due to altered land-use policies.

ACKNOWLEDGMENTS

Research supported by the National Science Foundation under Interagency Agreement 40-700-78 with the U.S. Department of Energy under contract W-7405-eng-26 with Union Carbide Corporation. Publication No. 1615. Environmental Sciences Division, Oak Ridge National Laboratory.

LITERATURE CITED

Anderson, S.H., and H.H. Shugart. 1974 Habitat selection of breeding birds in an East Tennessee deciduous forest. Ecology 55:828-837.

Barr, A.J., J.H. Goodnight, J.P. Sall, and J.T. Helwig. 1976. A User's Guide to SAS-76. 329 p. SAS Institute, Inc., Raleigh, N.C.

Conner, R.N., and C.S. Adkisson. 1976. Discriminant function analysis: a possible aid in determining the impact of forest management on woodpecker nesting habitat. Forest Science 22:122-127.

Cody, M.J. 1968. On methods of resource division in grassland bird communities. American Naturalist 102:107-147.

Colwell, R.K., and D.J. Futuyma. 1971. On the measurement of niche breadth and overlap. Ecology 52:567-576.

Cooley, W.W., and P.R. Lohnes. 1971. 364 p. Multivariate Data Analysis. John Wiley and Sons, New York, N.Y.

Dixon, W.J., editor. 1974. BMD Biomedical computer programs. 773 p. University of California Press, Berkeley, Calif.

Dueser, R.D., and H.H. Shugart. 1978. Microhabitats in forest floor small mammal fauna. Ecology 59:89-98.

Dueser, R.D., and H.H. Shugart. 1979. Niche pattern in a forest floor small mammal fauna. Ecology 60:108-118.

Fujii, K. 1969. Numerical taxonomy of ecological characteristics and the niche concept. Systematic Zoology 18:151-153.

Green, R.H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve mollusks of central Canada. Ecology 52:543-556.

Green, R.H. 1974. Multivariate niche analysis with temporally varying environmental factors. Ecology 55:73-83.

Harner, E.J., and R.C. Whitmore. 1977. Multivariate measures of niche overlap using discriminant analysis. Theoretical Population Biology 12:21-36.

Harris, V.T. 1952. An experimental study of habitat selection by prairie and forest races of the deermouse, Peromyscos maniculatus. 103 p. Contribution of Laboratory of Vertebrate Biology 56. Univiversity of Michigan Press, Ann Arbor, Mich.

Hespenheide, H.A. 1971. Flycatcher habitat selection in the eastern deciduous forest. Auk 88:61-74.

Horn, H.S. 1966. Measurement of overlap in comparative ecological studies. American Naturalist 100:419-424.

Hutchinson, G.E. 1944. Limnological studies in Connecticut. VII. A critical examination of the supposed relationship between phytoplankton periodicity and chemical changes in lake waters. Ecology 25:3-26. (see footnote 5, p. 20).

Hutchinson, G.E. 1957. Concluding remarks. Cold Spring Harbor Symposium of Quantitative Biology 22:415-427.

Hutchinson, G.E. 1965. The ecological theater and the evolutionary play. 134 p. Yale University Press, New Haven, Conn.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Klopfer, P. 1965. Behavioral aspects of habitat selection. A preliminary report on stereotypy in foliage preferences in birds. Wilson Bulletin 75:15-22.

Kohler, W. 1947. Gestalt psychology. 482 p. Liverright Publishing Corp., New York, N.Y.

Lennartz, M.R., and J.P. McClure. 1979. Estimating the extent of Red-cockaded woodpecker habitat in the Southeast. p.48-62. In Frayer, W.E., editor. Forest resource inventories: Proceedings of a workshop. [Fort Collins, Co., July 22-27, 1979] Department of Forest and Wood Science, Colorado State University. Unnumbered publication.

MacArthur, R.M. 1958. Population ecology of some warblers of northeastern coniferous forests. Ecology 39:599-619.

MacArthur, R.M., and E. Piauka. 1966. On optimal use of a patch environment. American Naturalist 100:603-609.

Maguire, B., Jr. 1972. A partial analysis of the niche. American Naturalist 101:515-523.

Martinka, R.R. 1972. Structural characteristics of Blue Grouse territories in southwestern Montana. Journal of Wildlife Management 36:498-510.

May, R.M. 1975. Some notes on estimating the competition matrix. Ecology 56:737-741.

McNaughton, S.J., and L.L. Wolf. 1972. Dominance and the niche in ecological systems. Science 167:131-139.

Pielou, E.L. 1972. Niche width and overlap: a method for measuring them. Ecology 53:687-692.

Rosenzweig, M.L. 1973. Habitat selection experiments with a pair of coexisting heteromyid rodent species. Ecology 54:111-117.

Rosenzweig, M.L. 1974. On the evolution of habitat selection. p. 401-404. In Proceedings of First International Congress on ecological structure, function and management of ecosystems. Centre Agriculture Publication and Documents, Waginingen, The Netherlands.

Rosenzweig. M.L., and J. Winakur. 1969. Population ecology of desert rodent communities: habitats and environmental complexity. Ecology 50:558-572.

Schroder, G.D., and M.L. Rosenzweig. 1975. Perturbation analysis of competition and overlap in habitat utilization between Dipodomys ordii and Dipodomys merriami. Oecologia 19:9-28.

Shugart, H.H., S.H. Anderson, and R.H. Strand. 1975. Dominant patterns of bird populations of the eastern deciduous forest biome. p. 90-95. In Smith, D.R., technical coordinator. Management of forest and range habitats for nongame birds: Proceedings of a symposium [Tucson, Ariz., May 6-9, 1975] USDA Forest Service General Technical Report WO-1, 343 p. Washington, D.C.

Shugart, H.H., R.D. Dueser, S.H. Anderson. 1974. Influence of habitat alterations on bird and small mammal populations. p. 92-96. In Slusher, J.P., and T.M. Hinckley, editors. Timber-wildlife management: Proceedings of a symposium [Columbia, Mo., Jan. 22-24, 1974] Missouri Academy of Science Occasional Paper 3, 131 p. Columbia, Mo.

Shugart, H.H., and B.C. Patten. 1972. Niche quantification and the concept of niche pattern. p. 283-327. In Patten, B.C., editor. Systems analysis and simulation in ecology. Volume 2. Academic Press, New York, N.Y.

Shugart, H.H., and D.C. West. 1977. Development of an Appalachian deciduous forest succession model and its application to assessment of the impact of the chestnut blight. Journal of Environmental Management 5:161-179.

Shugart, H.H., and D.C. West. 1980. Forest succession models. BioScience 30:308-313.

Smith, T.M., H.H. Shugart, and D.C. West. 1981. FORHAB: A forest simulation model to predict habitat structure for non-game bird studies. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Tinbergen, N. 1951. The study of instinct. 228 p. Oxford University Press, London.

Von Uexkull, J. 1909. Umwelt und Innenwelt der Tiere. Springer-Verlag, Berlin.

Wecker, S.C. 1963. The role of early experience in habitat selection by the prairie deermouse, Peromyscus maniculatus bairdi. Ecological Monographs 33:307-325.

Wiens, J.A. 1969. An approach to the study of ecological relationships among grassland birds. 93 p. Ornithological Monographs 8.

DISCUSSION

MICHAEL C.S. KINGSLEY: Multivariate analysis tends to assume multivariate normality. Are there considerations in niche theory, particularly with respect to niche packing, which imply that niches should be a) similar in shape and orientation, b) similarly oriented multivariate normal density spaces?

H.H. SHUGART: Yes. R.M. May (1973. Stability and complexity in model ecosystems. 265 p. Princeton University Press, Princeton, N.J.) used the standard deviation and separation of means of species distributions along a continuum as an index of packing and overlap. He then determined ratios of these two statistics from derivations based on a general n-species non-linear competition model. He (Chapter 6) also provided a fair number of citations on niche shapes. In a somewhat less abstract vein, the late R.H. Whittaker generally pictured the distributions of species in response to gradients as unimodel and of shapes that are easily approximated by normal distributions. This is true both in his data papers, as well as in his more theoretical works. Whittaker used fairly abundant literature citations and his work serves as a useful point of reference. Maguire (1972) studied niche shapes in protozoa using non-parametric methods and found niches to have tendencies to vary along similar environmental axes and to have similar shapes. R.H. MacArthur and E.P. Wilson (1967. The theory of island biogeography. 203 p. Princeton University Press, Princeton, N.J.) produced a "compression hypothesis" regarding the expected similarity in shape and/or orientation of niches in different communities. W.E. Westman (1980. Gaussian analysis: identifying environmental factors influencing bell-shaped species distributions. Ecology 61:733-739.) provides a discussion on shapes of species' responses to gradients and also provides several references to individuals who have noted normal distributions in nature. A discussion with several citations of theories that would lead to such shapes is found in Westman's introduction.

THE USE AND MISUSE OF STATISTICS IN WILDLIFE HABITAT STUDIES[1]

Douglas H. Johnson[2]

Abstract.--This paper briefly surveys the application of various multivariate statistical techniques in studies of wildlife and their habitats. Several methods are widely employed, but with little regard for the requisite assumptions and often without full appreciation of what the methods do and whether they are appropriate for the problem or not.

The well known fact that species typically respond to an environmental gradient in a nonlinear fashion is poorly accounted for in many statistical treatments. Moreover, most analyses are not well validated. The few valiant attempts at validation have suggested that the models produced, often for predictive or management purposes, are less successful than might have been anticipated.

The use of multivariate methods in developing recommendations for wildlife management calls for special caution, for it is a major step from describing the relationships observed between a species and some habitat features to predicting the response of that species as the habitat changes.

Key words: Canonical correlation analysis; discriminant function analysis; multiple regression; nonlinear response function; principal components analysis; transformations; validation.

INTRODUCTION

The application of multivariate analysis to studies of wildlife habitat offers an exciting opportunity to statisticians. They have a major role to play as wildlife studies become increasingly complex and as greater numbers of environmental variables are investigated. A basic tenet of ecology is that "everything is connected to everything else." Although that generalization is a bit extreme, it does express the ecologist's conviction that the variables affecting a species or a system are numerous.

Faced with this acknowledged need for statistical service and counsel, statisticians must respond appropriately, which means carefully and thoughtfully, with a full understanding of the biological problem, and with an appreciation of the consequences of the statistical analysis and interpretation. Merely adding variables to an analysis will not do. Multivariate analysis must not be a masquerade for ignorance.

In this paper I survey some of the potential applications of multivariate analysis to wildlife-habitat studies. I try to offer biologists some guidance through the maze of multivariate statistics. My viewpoint is that of

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop. April 23-25, 1980, Burlington, Vt.

[2]Statistician, U.S. Fish and Wildlife Service, Jamestown, ND 58401.

a statistician, with one eye on the mathematical requirements of the methodology, and one eye on the needs of the ultimate user--the wildlife manager. While looking in these two directions, we also must not forget the animal; its biology is of fundamental importance.

IS MULTIVARIATE ANALYSIS APPROPRIATE?

Green (1971) identified the need for a multivariate approach when he mentioned three operational problems in defining the niche of a species: 1) Not all potentially important environmental parameters can be measured. 2) Many of the parameters measured are likely to be correlated, relatively invariant, or irrelevant to the problem. 3) The many potentially relevant variables result in a mass of multidimensional data that is difficult to interpret.

These considerations have led ecologists to the troughs of principal component analysis and discriminant function analysis, where they have drunk freely. These analyses have produced "new" variables that are linear combinations of the old ones, and which are fewer in number, to eliminate the third problem, a large mass of data. Further, these new variables are uncorrelated, which overcomes the second problem. It is left to the ecologist to handle the first problem by carefully including variables that are potentially important to the species; statistics will be of minimal help here.

These multivariate analysis overcome two obstacles, but not without their own disadvantages. One drawback is that, by involving linear combinations of all the measured variables, they do not really reduce the parameter space; all the original variables must be measured in order to calculate the new ones. In that sense, the analyses only provide guidance for future research. Another problem is that the linear combinations themselves may be extremely difficult to interpret. Many of the published ones lend themselves to meaningful interpretation, although some do not; and I suspect that a lot of unintelligible linear combinations have been lost somewhere between analysis and publication.

I wonder if we can take a slightly different approach. Can we objectively define meaningful variables *a priori*, instead of doing so statistically, *a posteriori*? These new variables should be few in number, more or less uncorrelated with one another, and of potential importance to the animal. These variables would account for what is known (or believed) about the animal, and any additional variables that the analysis identified as significant would represent new findings worthy of additional investigation. For example, a bird might find James' (1971) outline drawings of niche-gestalts to be meaningful, and would be willing to select its habitat based on those drawings. But the bird would be hard-pressed to plug the values of 15 or 20 variables into a number of linear combinations, compare the calculated values to one another, and select a habitat with value closest to its liking.

Multivariate analysis, while evidently appropriate for wildlife-habitat studies, is difficult to apply and understand; Gnanadesikan (1977:2) suggested that univariate difficulties are raised to the *p*th power. He also identified some of the added problems and noted that much of the theoretical work in multivariate analysis, oriented toward formal procedures such as hypothesis tests on means, is of limited value for actual data analysis in the multivariate case. I next discuss some of the usual multivariate methods and their role in wildlife-habitat work, as identified by speakers in this workshop and earlier publications.

Discriminant Function Analysis

One of the most widely used multivariate techniques is discriminant function analysis (DFA), which is used to separate observations into groups on the basis of a set of measurements. In wildlife-habitat studies the observations are usually sites, and the groups denote whether a species was present or absent from the site, or whether species A was present as opposed to species B present. The variables are habitat measurements at the site. Lachenbruch (1975) is a valuable general reference and Tatsuoka (1970) presented a nontechnical exposition on DFA. Williams (1981) noted that DFA is applicable when the groups are well-defined and the set of measurements is ecologically meaningful. Groups of habitat sites defined by the presence or absence of a species are not always well-defined. At least three reasons for a species' absence can be identified: 1) the habitat at the site is unsuitable; 2) the habitat is suitable but the species is absent for other reasons, such as numbers in the population inadequate to occupy all suitable habitat or interspecific competition; or 3) the habitat is suitable and the site occupied, but the sampling procedure failed to detect it. DFA based on presence-absence data involves the implicit assumption that absence is due to the first reason above, but even in that situation the group corresponding to "species absent" may include sites where the habitat is unsuitable for different reasons (see subsequent section on Nonlinear Response Functions).

The requirement of DFA for ecologically meaningful measurements has already been addressed. My experience with biologists suggests that on the basis of their prior knowledge of a species and its habitats, they often can develop a very limited number of ecological variables that characterize occupied habitat. These variables tend to be combinations of two or more habitat measurements; for example, the breeding habitat of American woodcock (*Philohela minor*) might be characterized by a single variable expressing the presence of fertile soil supporting earthworms, vegetation dominated by shrubs or young trees, and the nearby presence of forest openings.

Ecologists should use their best information when defining habitat variables.

Williams (1981) emphasized the importance of equality of covariance matrices in DFA. More widespread is the notion, expressed by Klecka (1975:435) that DFA "is very robust and these assumptions (multivariate normality and equal covariance matrices) need not be strongly adhered to." Lachenbruch (1975) reviewed several studies and concluded that linear discriminant functions are satisfactory if the covariance matrices are not too different. If they differ considerably, then the appropriate technique is quadratic discrimination, which employs the unequal covariance matrices in the discriminant functions. Unfortunately, quadratic DFA requires large samples and is itself not robust to nonnormality. Most of the work involving DFA and its sensitivity to underlying assumptions has focused on the misclassification rates of the discriminant functions. In ecological studies the interpretation of coefficients is also important, but little is known of their robustness.

Principal Components Analysis

Principal components analysis (PCA) and closely related factor analysis are multivariate procedures designed to reduce the dimensionality of a data set. The purpose of the reduction might be to "stabilize scales of measurements" (Gnanadesikan 1977:6) by compounding several measurements of a similar nature into a fewer number that may be more stable, but I am aware of no ecologist who stated this as an intention. More often the purpose is for exploratory analysis, to screen out of many variables a few that are important. Bhattacharrya (1981) provided a general introduction to these methods; Gnanadesikan (1977) compared several methods including nonmetric and nonlinear ones.

Three problems specific to the application of principal components analysis in ecological work can be identified. First, I see little justification for selecting a linear combination of variables simply because it maximizes the variance within the total set, that is, it is a "best" summary. As Holmes et al. (1979) noted, one can increase the percentage of variance explained by the first principal component merely by adding redundant variables to the data set. As more and more of these are included, the principal components analysis appears better and better, but in fact only noise is being added to the system, and interpretation becomes increasingly awkward. A large percent of variance explained may reflect only ignorance in the selection of variables.

Second, PCA is not always useful in discovering the underlying structure of the variables. Armstrong (1967) presented an example involving 11 measures on rectangular boxes of various metals. All variables were functions of five underlying (and independent) factors: length, width, thickness, density and cost per pound. The PCA identified three factors that summarized 90.7 percent of the information in the original 11 variables. Despite an orthogonal rotation, the factors were not easy to interpret and the PCA was unable to reveal the rather simple basic structure of the data. It would seem unreasonably sanguine to expect PCA to be more enlightening in a complex ecological system. Karr and Martin (1981) presented another example, this one pertinent to biological data, that illustrated some hazards associated with the use and interpretation of PCA.

Third, there is no reason to assume that a principal component necessarily relates to the animal or its needs. The animal could be responding to one of the variables that is a minor component of all those that were measured. It seems more appropriate in most circumstances to use regression analysis with some measure of population density or "fitness" as the dependent variable. Smith (1981) suggested that canonical correlation analysis may also be useful in this regard.

In general, PCA offers a convenient way of summarizing a data set containing many variables. If the ecologist's purpose in measuring those many variables is to relate them to the presence/absence or density of a wildlife species, it is not obvious that the major principal components are the important ones. If most variables were selected because of prior knowledge about their relationship to the species, then these variables are likely to appear in the major components. It is conceivable, however, that a minor component may be important to the animal, and for this reason it is suggested that the species' response be examined in relation to each of the components. Factor analysis, which essentially deletes minor components, would not permit this examination and should thus be avoided in initial analyses.

Canonical Correlation Analysis

Canonical correlation analysis involves the linear relationship of one set of variables to another; it has had limited application to ecological problems. Smith (1981) reviewed these applications and discussed the technique and its shortcomings. He also provided a useful introduction to the literature. My own belief is that canonical correlation will continue to play only a small role in habitat studies. In addition to the reasons given by Smith, viz., difficulties in interpretation, large sample size requirements, lack of robustness to nonnormality, and assumption of linearity, the rationale for the technique rests upon the ecologist showing concern for a linear combination of "dependent" variables. Moreover, the coefficients in that combination are not based upon the biologist's intuition or interests, but are generated by the technique. How does one justify an interest in 2.07 robins + 0.16 brown thrashers - 1.92 chickadees? More likely the ecologist has a single dependent variable, or set of them, in mind, and univariate

or multivariate regression is the more appropriate method.

Other Techniques

Regression has not been specifically addressed in this workshop, although it remains one of the most popular and powerful statistical tools for examining relationships among variables. The usual formulation of the regression model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \text{error},$$

where the X's are fixed input (or "independent") explanatory variables, is not strictly a multivariate technique; there is only one random variable, the error term. In actual practice, however, the X's are not fixed quantities (Johnson 1981) and the model can be viewed profitably in a multivariate light. If several species are to be examined in relation to a set of habitat variables, multivariate regression appears to be the appropriate method. Gnanadesikan (1977) discussed the method and noted (p. 81) that a multivariate viewpoint may be preferable to considering each dependent variable separately, because of intercorrelations among the dependent variables. Unfortunately, most multivariate regression work has focused on hypothesis testing (the general linear model) rather than model building.

In contrast to their botanical counterparts, wildlife ecologists have made little use of clustering techniques. Cluster analysis differs from DFA (a classification technique) by the lack of groups defined a priori; the groups are essentially defined by the clustering algorithm and the user is responsible for providing a reasonable interpretation of the resulting groups. Cluster analysis differs from ordination in the implicit assumption that units fit into neat and discrete groups, rather than being arranged in a continuum along one or more principal component or discriminant function axes.

Although multivariate techniques can be and have been rewarding in habitat studies, ecologists should not ignore the simple and powerful methods of univariate statistics. Careful analyses begin with graphical displays of the data for several purposes: detecting outlying observations that may be erroneous or would be inordinately influential on the analysis; selecting variables that appear to be important; checking the assumptions underlying particular analyses, such as normality; and suggesting appropriate transformations of the variables. Valuable references on graphical techniques include Daniel and Wood (1971), who emphasized the use of plots for detecting relationships between variables and evaluating fitted models; Mosteller and Tukey (1977), who employed graphs both for general display and for fitting of models; and Green (1979), who discussed and illustrated many graphs useful in ecological studies. Plotting is more awkward in the multivariate situation than in the univariate case, but methods are available (e.g., Gnanadesikan 1977). Among these are 1) two- and three-dimensional scatterplots of subsets of the data for studying separation within the sample, outliers, general shape and interrelationships; 2) probability plots of the observations on each response, useful for suggesting transformations; 3) scatterplots or probability plots onto principal components, and others. The longest journey begins with a single step. To insure that the direction of the trip is appropriate, that first step should be graphing the data.

Transformations

It is often necessary to transform variables in order to meet more closely the assumptions of various statistical methods, specifically those of normality, constant error variance, and uncorrelated errors. Other purposes are independent of the statistical properties, such as simplifying the relationship between variables or quantifying qualitative, count, or percentage data. Transformations may also maximize the separation between groups of observations.

Univariate transformations have received much more attention than multivariate ones; good expositions are given by Green (1979:43-54), Kruskal (1978:1044-1056), and references they cited. Kruskal (1978) offered some "clues" that suggest a transformation might be in order. Variables that approach an intrinsic boundary may be candidates for transformation; examples are percentages that closely approach 0 or 100 percent or correlation coefficients that come close to ± 1. A nonnormal distribution can be detected by plotting observations on normal probability paper or by graphical procedures available in SAS (SAS Institute Inc. 1979) and other statistical packages. Nonconstancy of variance can be statistically tested for, but graphical procedures are more informative. Plots of residuals from fitted models may be made to suggest the presence of correlated error terms. Plots not only indicate the need for transformations, they also point out possible outlying observations. These may be data errors or at least data points more influential than the others.

The choice of an appropriate transformation, once one is recognized as necessary, is not quite as confusing as it might seem from the variety of those available. In many situations the choice can be made from theoretical considerations, such as the arcsine transformation for binomial data. In other situations a suitable selection can be made by examining the data; for example, the well known Box and Cox (1964) power transformation is a family of transformations indexed by a parameter that is itself estimated from the data.

Less guidance is available for transformations of multivariate data sets. This is unfortunate, because multivariate situations may require transformations more often in order to use familiar and simple analytic methods (Machado

1976). The choice of models in multivariate situations is more limited due to the general dependence on strict normality. The simplest approach to multivariate transformation is to treat each variable separately; this is likely to be a good first step at least. As Dunn (1981) noted for normalizing transformations, marginal normality does not insure multivariate normality, but the exercise is harmless at worst and is often quite adequate. Andrews et al. (1971) and Machado (1976) generalized the Box and Cox transformation to the multivariate case, and Dunn (1981) extended it to cover situations with more than one sample.

Transformations are often useful for obtaining stricter compliance with certain assumptions of statistical analysis. My own experience has demonstrated little advantage from analysis of transformed, as opposed to untransformed, data. Careful design to insure randomness overcomes many problems with correlated errors, and reasonably large and well-balanced samples mitigate the effects of nonnormality and nonhomogeneous variance (Green 1979:165). Transformations are advised, however, if they will clarify relationships or if the statistical assumptions are obviously violated. It certainly will not hurt to analyze the data both untransformed and transformed, and judge which analysis is superior.

NONLINEAR RESPONSE FUNCTIONS

It seems generally agreed that a species responds to an environmental variable, or to a gradient, in a nonlinear fashion. The form of the function could be normal, or of another symmetric shape, or it could be asymmetrical but unimodal, or it could even be bimodal. But it is nonlinear. Despite this wide acknowledgment, relatively little has been said about the effect of nonlinearity on the results of multivariate analysis (Westman 1980).

I wonder if we are like the several blind men, each touching a different part of an elephant, and reaching widely discrepant conclusions about the total shape of the animal. Suppose, for example, that a species responds to an environmental variable X as shown by the symmetric curve in figure 1. An investigator studying values of the variable only in region A would conclude that the species responds favorably to the variable; it is an "increaser." A study in region C would lead to just the opposite conclusion, while a study in region B would probably find that the species did not correlate at all with the variable. All conclusions are correct, and yet all are wrong. Even a study involving the full range of the variable, regions A plus B plus C, would detect no linear association.

A possible example is that of Converse and Morzuch (1981), who found that hare activity was correlated positively with the number of hardwood trees on one of their study areas, while the correlation with hardwood trees was negative in the other area.

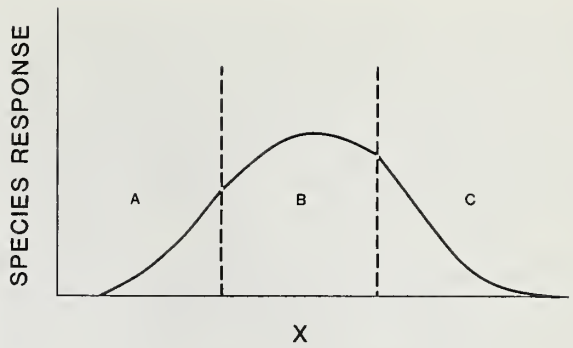


Figure 1. The nonlinear response of a species to an environmental gradient X.

In regard to discriminant analysis, Green (1971:544) suggested that a nonlinear response function might not be troublesome, because "species are groups separated in ecological space by linear additive functions of ecological parameters, rather than dependent variables supposedly related to the ecological parameters in a linear additive manner." I am not convinced that there is no difficulty, at least in some applications. Suppose, referring back to the response curve (fig. 1), that we form two groups: species Y present, and species Y absent. We want to distinguish these groups on the basis of their value of X. Most X values in regions A and C would fall into the species absent group, while region B would donate mostly species present. But it is evident that average values of X in the two groups might be very similar, even identical. For example, if the species is a mesic one and X represents a moisture gradient, then some sites might be too dry and others too wet. There are thus two groups in which the species is absent, and they differ more from each other than they do from the species present group. In higher-dimension space the problem only gets worse. Quadratic discriminant functions are theoretically applicable to such situations, but separating groups on the basis of variance when means are similar is tenuous at best.

In principal components analysis, or factor analysis, two problems related to nonlinearity occur. First, the methods produce linear combinations of the variables; if a nonlinear function is a better summary of the data, we will not detect it by the usual methods, although some nonlinear techniques have been developed (Gnanadesikan 1977). Second, if a species responds nonlinearly to a principal component, we had better be alert if we want to detect it, e.g., by plotting species responses against each component.

Nonlinear response functions can be detected by graphical tools and can be treated appropriately by nonlinear regression models (cf.

Boyce [1981], who eliminated from analysis a variable with a recognized nonlinear effect).

We must be particularly careful about nonlinearity when presenting our results to resource managers, who look assiduously for the "bottom line": What can a manager do? He might be told that opening up a closed forest yields more woodpeckers. He also needs to be told that this practice works only up to a point: complete elimination of trees will certainly have the opposite effect. We must remind him, and ourselves, that a management activity is like aspirin: just because two pills are good for us, it does not follow that four are twice as good, and a whole bottle is ideal.

STATISTICAL ASSESSMENT

I now turn to the question of how we might assess the statistical properties of habitat analyses. As a statistician, I am frequently appalled by the small sample sizes reported in many studies, particularly those involving large numbers of variables. Small samples of a complex system cannot support sound conclusions.

We are taught early in our statistical training to look, not only at the mean of a distribution, but also at its variability. In many published studies, however, we see species means plotted on axes of principal components or discriminant functions. Relatively few investigators mention the variation about those means; the few who have done so were very informative.

Conner and Adkisson (1977) plotted means of principal components for five species of woodpeckers; niche separation was shown among the species. When values for individual birds were plotted, however, they exhibited more overlap than might have been expected. Raphael (1981) also found considerable overlap among species on discriminant axes. Similarly, Smith's (1977) analysis of summer birds along a moisture gradient showed clear separation of 1% confidence regions for the means of eight species on two principal component axes. When plotted on the discriminant function representing a moisture gradient, however, values for individual birds of five species were scattered throughout, and relatively minor separation was evident. Researchers should meticulously examine the variability among animals.

Too little attention is given to annual variation in bird populations, fluctuations that take place seemingly without regard to habitat conditions, and certainly without regard to the unfortunate ecologist who is trying to determine relationships between habitats and populations. Ornithologists familiar with long-term studies can document the magnitude of this variation (e.g., Lack 1966, Wiens 1975), and I strongly suspect that it impacts habitat analysis. Rice et al. (1981) illustrated this problem nicely; discriminant functions developed from one year's data had few successes at predictions during the next year.

The use of stepwise procedures is viewed by many statisticians as a "fishing expedition." The standard significance tests are invalid (Draper et al. 1971, Pope and Webster 1972, Rencher and Larson 1980) and results are questionable, particularly in the usual situation of numerous variables. Automated computer procedures are no substitute for careful biological reasoning.

I believe that the application of robust statistical methods will offer considerable help in habitat analysis. McDonald (1981) gave details and references to the literature. Harner and Whitmore (1981) exemplified these methods and some internal validation tools, such as the jackknife and leaving-one-out procedures. In multivariate situations, applications of robust methods thus far have been primarily the treatment of one variable at a time (Gnanadesikan 1977:136), but this is likely to change in the near future. Particularly appealing is robustness of result with respect to method. If the same general conclusions are reached through the use of several different methods, each resting upon a somewhat different set of assumptions, we should feel more comfortable about the validity of those conclusions (Green 1977:14).

Validation is an important final product of an analysis. Whitmore (1977:263) stated that "the validity of ordination work can be tested by subsequent field observation." I agree, but such testing is rarely done. Noon and Able (1978) applied their discriminant functions for thrushes, developed for five species on Mount Mansfield in Vermont, to the two species occurring in the Great Smoky Mountains in Tennessee and North Carolina. They found little predictive ability.

The validation of models is particularly important if we are to present them to resource managers for their use. The need for caution on our part is obvious. In statistical parlance, management is really control, which is farther up the methodological ladder than description, inference, and prediction. As ecologists, our abilities on the lower rungs are as yet unproven.

A question that is rarely raised: What is the universe to which our results are to pertain? If it is a single study area, in a single year, with the measurements we have observed, there is no problem. If we want to generalize, let us be careful. Our study area must be representative of the area we want to extrapolate to, similarly the year, and the habitat. Noon and Able (1978) and Converse and Morzuch (1981) described how naive application of results from one area to another could be erroneous.

I see a clear need for experimentation. As observers of (more-or-less) natural populations of wild animals that do as they please, our options

here are unforunately limited. But they are not curtailed. Is there a role for removal studies such as Stewart and Aldrich (1951) performed? An area which was repopulated soon after birds were removed would exhibit more of whatever the birds need than would an area that remained vacant. Habitat manipulation is another possibility. Some imaginative thinking might produce very valuable experiments to test the results we obtain from multivariate analysis of passive and uncontrolled observations.

MANAGEMENT APPLICATIONS

Multivariate analysis has been suggested as an important and useful tool in the management of habitats (e.g., Lennartz and Bjugstad 1975, Conner and Adkisson 1976, Evans 1978, Noon and Able 1978, Niemi and Pfannmuller 1979). It promises enhanced ability to examine a multitude of variables at once, just as a resource manager manipulates a multitude of habitat variables with the swing of an ax or the drop of a match. Clear-cutting a forest does not merely change the standing biomass of trees, it also affects stem counts, ground cover, litter, and a host of other measurable features: clearly a multivariate situation.

Again, I must sound a cautionary note. Are we looking at the right variables, and are we confusing association with causation? Consider the following paradigm. An action A, which may be either a natural phenomenon or a specific management activity, causes certain effects, denoted X_1, X_2, ..., X_n, upon the environment. One of these environmental effects, say X_1, in turn triggers a population response by a particular species, Z. Suppose we are observing the phenomenon, and record many of the environmental variables, but not X_1, and we record the response of species Z. Our analysis would show a relationship between the X's and Z; a multivariate analysis might reduce the set of X's to a smaller set of principal components, or discriminant functions, that also relate to Z. In truth, however, these associations are all spurious; the real association involves the "lurking variable" X_1, which correlates with the other X's and causes the response by species Z.

Even if we are clever enough to measure X_1 along with the other X's, the true relationship between X_1 and Z is almost certain to be clouded by the plethora of other variables. This paradigm is certainly not new; we have all been exposed to it in one form or another. But I believe we need to remind ourselves of it regularly; I am deeply concerned about the soundness of management recommendations based upon associations that are interpreted as causations.

The analysis may not always be misleading. In the paradigm just presented we might conclude that a particular configuration of X values is conducive to good numbers of species Z. The way to reach that X-configuration is by applying action A. This advice will work. Action A affects variable X_1, which in turn results in an increase of species Z.

But suppose there are other ways to obtain the desired configuration of X values without the appropriate value of X_1. Then that management action, even though it succeeds in producing the "correct-appearing" habitat, will fail to increase species Z.

As an example, fire produces certain effects in North Dakota grassland habitats, including removal of most standing vegetation and reduction of litter. The effects of the fire in turn produce a response by shorebirds; killdeer (Charadrius vociferus), marbled godwits (Limosa fedoa) and upland sandpipers (Bartramia longicauda) come to the burned areas to forage. Some range ecologists claim that appropriate regimens of cattle grazing will produce habitat changes similar to those caused by fire. Grazing could in fact produce similar measurements on a variety of vegetative parameters. But the shorebirds do not seem to respond in the same way to grazing as they do to fire. They must be keyed into one or more of the "lurking variables."

Other examples might contrast the effects of forest fires to those of clear-cutting, natural food supplies to artificial feeding stations, or natural pest control versus chemical control.

Rice et al. (1981) also pointed out how management practices directed toward one season can have possibly unintended and undesirable ramifications during other parts of the year.

Where we cannot design experiments freely, select optimal levels of variables at will, and replicate as often as we desire, we must be reserved about our findings. If management practices are adopted as a result of a multivariate habitat analysis, a thorough evaluation should be designed and conducted to determine if predictions from the proposed model are accurate and the model thereby remains tenable.

CONCLUSIONS

Multivariate analysis has been claimed to be useful in studies of wildlife and their habitats, despite the fact, clearly pointed out in this workshop, that the assumptions of the various methods are largely unmet. I see two possible explanations for this seeming inconsistency. The first is that biologists might be reporting the results of sophisticated multivariate techniques only when they are in accord with their biological

intuition, previous knowledge, or results of simpler univariate analyses. A technique that produces conclusions in conflict with other knowledge may be rejected as inappropriate. If this is the case, then the multivariate methods are not truly useful; they only lend an appearance of further credibility to conclusions already reached from other directions.

A second possibility is that the multivariate methods might actually be more robust than is recognized. They may yield generally correct answers despite rather flagrant violations of their assumptions. Statisticians could be looking at finer detail than biologists do when they evaluate multivariate methods. A technique can be biased, of low power, inefficient, and defective in other statistical ways, but still provide valuable insight into the biological problem. One might define a technique to be useful if it produces more nearly correct answers than would be available without it.

It is for statisticians and biologists to work together in determining how useful various techniques are, relaxing assumptions that may not be critical, and designing studies to meet the assumptions that are truly needed. The papers and discussions have given a real stimulus to further cooperative efforts, and they will furnish guidance to researchers for many years to come.

ACKNOWLEDGMENTS

I appreciate comments on an earlier draft of this report supplied by D. E. Capen, J. C. Lewis, S. G. Machado and G. J. Niemi.

LITERATURE CITED

Andrews, D.F., R. Gnanadesikan, and J.L. Warner. 1971. Transformations of multivariate data. Biometrics 27:825-840.

Armstrong, J.S. 1967. Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. American Statistician 21:17-21.

Bhattacharrya, H. 1981. Theory and methods of factor analysis and principal components. Proceedings of this workshop.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society B 26:211-252.

Boyce, M.S. 1980. Robust canonical correlation of sage grouse habitat. Proceedings of this workshop.

Conner, R.N., and C.S. Adkisson. 1976. Discriminant function analysis: a possible aid in determining the impact of forest management on woodpecker nesting habitat. Forest Science 22:122-127.

Conner, R.N., and C.S. Adkisson. 1977. Principal component analysis of woodpecker nesting habitat. Wilson Bulletin 89:122-129.

Converse, K.A., and B.J. Morzuch. 1981. A descriptive model of snowshoe hare habitat. Proceedings of this workshop.

Daniel, C., and F.S. Wood. 1971. Fitting equations to data. 342 p. John Wiley and Sons, New York, N.Y.

Draper, N.R., I. Guttman, and H. Kanemasu. 1971. The distribution of certain regression statistics. Biometrika 58:295-298.

Dunn, J.E. 1981. Data-based transformations in multivariate analysis. Proceedings of this workshop.

Evans, K.E. 1978. Forest management opportunities for songbirds. Transactions North American Wildlife and Natural Resources Conference 43:69-77.

Gnanadesikan, R. 1977. Methods for statistical data analysis of multivariate observations. 311 p. John Wiley and Sons, New York, N.Y.

Green, R.H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of central Canada. Ecology 52:543-556.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. 257 p. John Wiley and Sons, New York, N.Y.

Harner, E.J., and R.C. Whitmore. 1981. Robust principal component and discriminant analyses of two grassland bird species' habitat. Proceedings of this workshop.

Holmes, R.T., R.E. Bonney, Jr., and S.W. Pacala. 1979. Guild structure of the Hubbard Brook bird community: a multivariate approach. Ecology 60:512-520.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Johnson, D.H. 1981. How to measure habitat--a statistical perspective. Proceedings of this workshop.

Karr, J.R., and T.E. Martin. 1981. Random numbers and principal components: further searches for the unicorn. Proceeding of this workshop.

Klecka, W.R. 1975. Discriminant analysis. p. 434-467. In Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Bent, editors. SPSS: statistical package for the social sciences. Second edition. McGraw-Hill Book Co., New York, N.Y.

Kruskal, J.B. 1978. Transformations of data. p. 1044-1056. In Kruskal, W.H., and J.M. Tanur, editors. International encyclopedia of statistics. Volume 2. The Free Press, New York, N.Y.

Lachenbruch, P.A. 1975. Discriminant analysis. 128 p. Hafner Press, New York, N.Y.

Lack, D. 1966. Population studies of birds. 341 p. Clarendon Press, Oxford.

Lennartz, M.R., and A.J. Bjugstad. 1975. Information needs to manage forest and range habitats for nongame birds. p. 328-333. In Smith, D.R., technical coordinator. Management of forest and range habitats for nongame birds: Proceedings of a symposium [Tucson, Ariz., May 6-9, 1975]. USDA Forest Service General Technical Report WO-1, 343 p. Washington, D.C.

Machado, S.B.G. 1976. Transformations of multivariate data and tests for multivariate normality. Ph.D. Dissertation. 279 p. University of Chicago, Chicago.

McDonald, L. 1981. A discussion of robust procedures in multivariate analysis. Proceedings of this workshop.

Mosteller, F., and J.W. Tukey. 1977. Data analysis and regression. 588 p. Addison-Wesley Publishing Co., Reading, Mass.

Niemi, G.J., and L. Pfannmuller. 1979. Avian communities: approaches to describing their habitat associations. p. 154-178. In DeGraaf, R.M., and K.E. Evans, compilers. Management of north central and northeastern forests for nongame birds: Proceedings of a workshop [Minneapolis, Minn., January 23-25, 1979]. USDA Forest Service General Technical Report NC-51, 268 p. North Central Forest Experiment Station, St. Paul, Minn.

Noon, B.R., and K.P. Able. 1978. A comparison of avian community structure in the northern and southern Appalachian Mountains. p. 98-117. In DeGraaf, R.M., technical coordinator. Management of southern forests for nongame birds: Proceedings of a workshop [Atlanta, Ga., January 24-26, 1978]. USDA Forest Service General Technical Report SE-14, 176 p. Southeastern Forest Experiment Station, Asheville, N.C.

Pope, P.T., and J.T. Webster. 1972. The use of an F-statistic in stepwise regression procedures. Technometrics 14:327-340.

Raphael, M.G. 1981. Interspecific differences in nesting habitat of sympatric woodpeckers and nuthatches. Proceedings of this workshop.

Rencher, A.C., and S.F. Larson. 1980. Bias in Wilk's Λ in stepwise discriminant analysis. Technometrics 22:349-356.

Rice, J., R.D. Ohmart, and B.W. Anderson. 1981. Bird community use of riparian habitats: the importance of temporal scale in interpreting discriminant analysis. Proceedings of this workshop.

SAS Institute, Inc. 1979. SAS user's guide. 1979 edition. 494 p. Raleigh, N.C.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58:810-819.

Smith, K.G. 1981. Canonical correlation analysis and its use in wildlife habitat studies. Proceedings of this workshop.

Stewart, R.E., and J.W. Aldrich. 1951. Removal and repopulation of breeding birds in a spruce-fir forest community. Auk 68:471-482.

Tatsuoka, M.M. 1970. Discriminant analysis: the study of group differences. 57 p. Institute for Personality and Ability Testing, Champaign, Ill.

Westman, W.E. 1980. Gaussian analysis: identifying environmental factors influencing bell-shaped species distributions. Ecology 61:733-739.

Whitmore, R.C. 1977. Habitat partitioning in a community of passerine birds. Wilson Bulletin 89:253-265.

Wiens, J.A. 1975. Avian communities, energetics, and functions in coniferous forest habitats. p. 226-265. In Smith, D.R., technical cooordinator. Management of forest and range habitats for nongame birds: Proceedings of a symposium [Tucson, Ariz., May 6-9, 1975]. USDA Forest Service General Technical Report WO-1, 343 p. Washington, D.C.

Williams, B.K. 1981. Discriminant analysis in wildlife research: theory and applications. Proceedings of this workshop.

DISCUSSION

E. JAMES HARNER: I disagree that multivariate analyses are almost always just exploratory. The jackknife and bootstrap techniques offer rather robust ways to make statistical inferences.

DOUGLAS JOHNSON: Multivariate techniques certainly can be used for inferential or, as I termed it, confirmatory research. As they are usually employed in wildlife-habitat studies, however, their proper role is exploratory.

JAKE RICE: From your figure it seemed that you pointed out that nonlinearity was a major problem in PCA, and you were talking largely about nonlinear responses of species (say abundances) to the environmental gradient represented by the principal component. Does not that sort of nonlinearily show up by simply graphing abundance against PCA scores? In that sense nonlinear responses are no more problem than in any univariate study where a species may also have a nonlinear response to any measure of an environmental attribute, and nonlinear regression methods are readily available. Is not the nonlinearity problem in PCA (and other multivariate techniques) one that the variables used in the PCA (say several environmental or habitat measures) may be interrelated in nonlinear ways?

DOUGLAS JOHNSON: You are correct on all counts. Nonlinearity of response is a potential problem whether the technique be univariate or multivariate. Nonlinearity can be examined by simple graphing, as you suggest and as I heartily recommend. A further difficulty with multivariate techniques, particularly principal components analysis, is nonlinear relationships among the explanatory (in our case, habitat) variables. If the relationships are present and strongly nonlinear, then the best linear combination of habitat variables may not be very good at all.

CHARLES SMITH: What is meant by "robustness"?

DOUGLAS JOHNSON: Robustness generally means insensitivity to violation of the assumptions of a technique. It has been further refined (Mallows, C.L. 1979. Robust methods--some examples of their use. American Statistician 33:179-184) to incorporate three concepts: 1) resistance--insensitivity to a moderate number of bad values and to inadequacies in the model; 2) smoothness--a characteristic of a technique in that it responds only gradually to the introduction of a few errors; and 3) breadth--the extent to which a technique can be applied in a wide variety of circumstances. A technique is called robust if it yields at least approximately correct answers despite having its assumptions not fully met.

RANDOM NUMBERS AND PRINCIPAL COMPONENTS: FURTHER SEARCHES FOR THE UNICORN?[1]

James R. Karr[2] and Thomas E. Martin[3]

Abstract.--Analysis of biological data with an array of multivariate procedures has increased in recent years. While these powerful tools have considerable potential for producing more rigorous biological conclusions, they are subject to misuse. We have analyzed real biological data and random number matrices using principal components analysis (PCA) and have shown that: 1) percent of variation accounted for may be similar for both, especially for the second and higher principal axes; 2) loadings of the original variables on principal axes are often as high as those for real data; and 3) matrix size is an important determinant of amount of variation extracted by PCA. The relevance of these and other points is discussed in light of the need for more objective grounds for the interpretation of PCA.

Key words: Bird/habitat relationships; principal components analysis; random numbers; sphericity test.

INTRODUCTION

This volume on multivariate statistics and wildlife habitat is the most recent in a series of publications demonstrating the power and value of multivariate procedures in analysis of ecological data. But, like so many other tools and/or dogmas in the ecological sciences, we feel it is necessary to reflect a bit on the potential for misuse of these "new" procedures. We urge caution in the use of one of the hottest items in the arsenal of multivariate procedures--principal components analysis (PCA).

Many researchers have used this procedure in the past decade for its major intended purpose--to reduce the complexity of multivariate data to a more manageable set of compound variables. This is not without some danger as no testing procedures are commonly used to allow measurement or evaluation of the significance of results generated by use of principal components analysis. Interpretation of pattern has commonly been based on relatively subjective grounds.

TWO DATA SETS

Our objective was to compare results of PCA on a matrix of real biological data with results from analysis of a random number table of the same dimensions. The size of data matrices used for analysis of bird/habitat relationships varies significantly among studies. For illustrative purposes we selected a matrix size of 10 x 24 because of an early use of PCA in bird/habitat literature (10 vegetation variables and 24 bird species). (Throughout this paper we refrain from citing specific studies. The questions we raise here are directed generally rather than at any particular study.)

[1] Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2] Professor, Department of Ecology, Ethology, and Evolution, University of Illinois, 606 E. Healey, Champaign, IL 61820.

[3] Ph.D. candidate, Department of Ecology, Ethology, and Evolution, University of Illinois, 606 E. Healey, Champaign, IL 61820.

Table 1. Percent of variation accounted for by the first three principal components for two 10 x 24 matrices.

Matrix	Percent of variation by component			
	I	II	III	Cumulative
10 vegetation variables, 24 bird species	57	16	12	85
10 x 24 matrix of random numbers	25	14	14	53

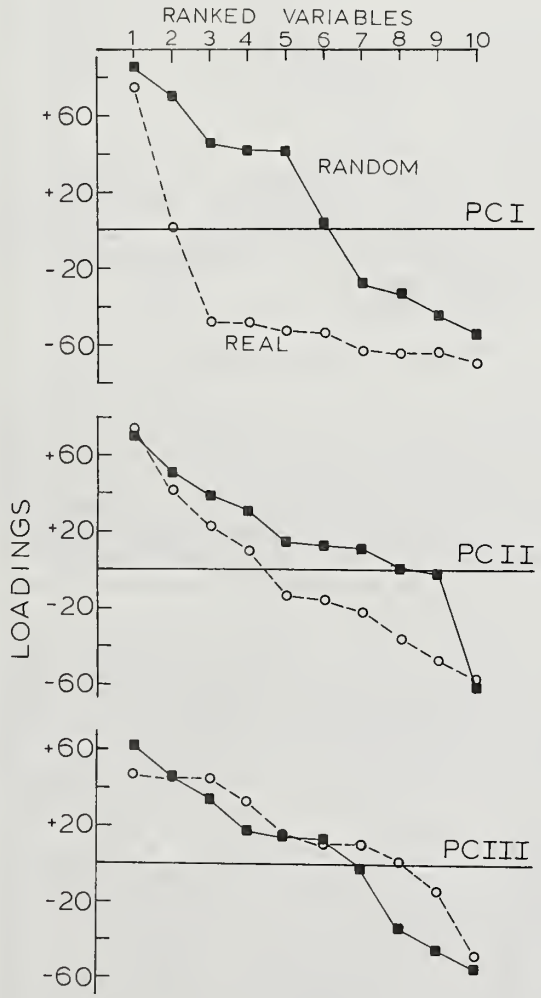


Figure 1. Distributions of loadings of original variables for principle axes I, II, and III for real biological data and random numbers (10 x 24 matrix).

For the real data matrix, 85% of the total variance was accounted for by the first three principal axes (table 1). However, use of PCA on an equivalent-sized matrix of random numbers extracted 53% of the variation in the first three components. Amount of variation accounted for in the real data was clearly very high, but it is instructive to view the situation from the opposite perspective. The percent of variation accounted for in a random number table was well above that which might be expected by the naive reader. A greater percentage of variation was accounted for by the first component for real data than for random numbers, but the second and third components accounted for similar amounts of variation in both real and random number matrices (table 1).

These results might tempt the reader to argue that loadings of original variables on the principal components must be much lower in the random numbers than the real data. This, it turned out, was not the case (fig. 1). Factor loadings had approximately the same maxima in the two matrices. We believe that most biologists would be able to generate a plausible post facto biological explanation of the high loadings in the random numbers for each of the three principal axes.

Although the maximum loadings of the original variables were about the same across the three principal axes for random and real data, the array of loadings over the ten variables was different for the first axis (fig. 1). Loadings for the random number and real data matrices were more similar in subsequent principal components. These results were due presumably to correlations among variables. The correlations among variables for real data include a greater number of high correlations than for random numbers (fig. 2). These high correlations allow the first component to account for a greater percentage of variation

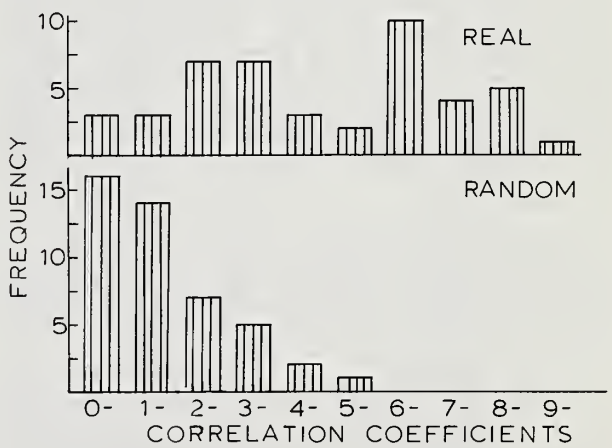


Figure 2. Distribution of correlations among original variables for real biological data and random numbers (10 x 24 matrix).

Table 2. Percent of variation accounted for by the first three principal components for real and random number (in parentheses) matrices of the same dimensionality. The real data examples are from published studies using principal components analysis.

Matrix size	Percent of variation by component			
	I	II	III	Cumulative
6 x 56	33(23)	22(20)	18(18)	73(61)
7 x 5	45(44)	23(28)	11(18)	79(91)
8 x 21	46(27)	17(20)	10(17)	73(64)
10 x 24	57(25)	16(14)	12(14)	85(53)
10 x 46	65(18)	12(16)	12(16)	85(50)
15 x 15	--*(22)	--(16)	11(13)	58(51)

* (--) indicates not available in original publication.

in real data. Thus, real data are more ellipsoidal in multivariate space than are random data which tend to be more spherical.

How, then, can we develop some confidence in the typically post facto explanations generated when real data are analyzed? As a first step, we suggest that authors and readers carefully interpret the percent of variation accounted for in a PCA by comparing the results with those obtained from a random number matrix of equal dimensions. One possibility might be a table of expected amount of variation accounted for by random number matrices that could be used as a test relative to the amount of variation accounted for in real data. Another possibility is the sphericity test described by Pimentel (1979). This procedure tests the equality of components. When three or more axes are of similar length, they define a sphere in which axes are arbitrarily placed. Thus, components may be uninterpretable when they are equal. Finally, there are other tests available when analysis is based on the covariance matrix. For instance, the equality of a vector based on biological data and a theoretically derived vector can be tested (Pimentel 1979). However, since most biological PCA's have been based on correlation matrices, we confine discussion to these analyses.

The sphericity test simply tests the homogeneity of the last p - k factors, where p is the total number of factors and k is the first k factors not being compared. The homogeneity of the second and third factors can be tested if the first three factors are transformed by subtracting the mean eigenvalue of the last p - 3 factors. Analysis of results from the 10 x 24 matrix (table 1) with the sphericity test shows that the second and third axes are equal for real data (χ^2 = 0.604, P > 0.05). Thus, real data appear to exhibit a primary ellipsoidal pattern (axis 1) but not in subsequent axes, so interpretation of axes 2 and 3 may be invalid.

OTHER MATRICES OF BIOLOGICAL DATA

We selected several published studies to illustrate the similarity of results from real and random number data. Much variation is explained by random number matrices relative to real data of the same dimensions, especially in situations with few variables (table 2). These results lead us to urge caution in using PCA for small matrices, a proviso on sample size similar to that for normal univariate statistical procedures.

As in the earlier example, the trend for a greater difference in the first axis but similar results for the second and third axes of random and real numbers seems to apply across a variety of matrix sizes (table 2). However, note that although the percent of variation accounted for in the real and random data is similar for the second and third axes, this is partly due to the first component of real data accounting for a greater percent of variation than in random numbers. This leaves less variation to be accounted for by the second and third components in real data. Real data account for a greater proportion of the remaining variation than in random data for large matrix sizes. For example, in the 10 x 46 matrix, the proportion of remaining variation accounted for by the second (real - 34%; random - 20%) and third (real - 52%; random - 24%) components is higher for real data. However, for the 7 x 5 matrix, the proportion of remaining variation

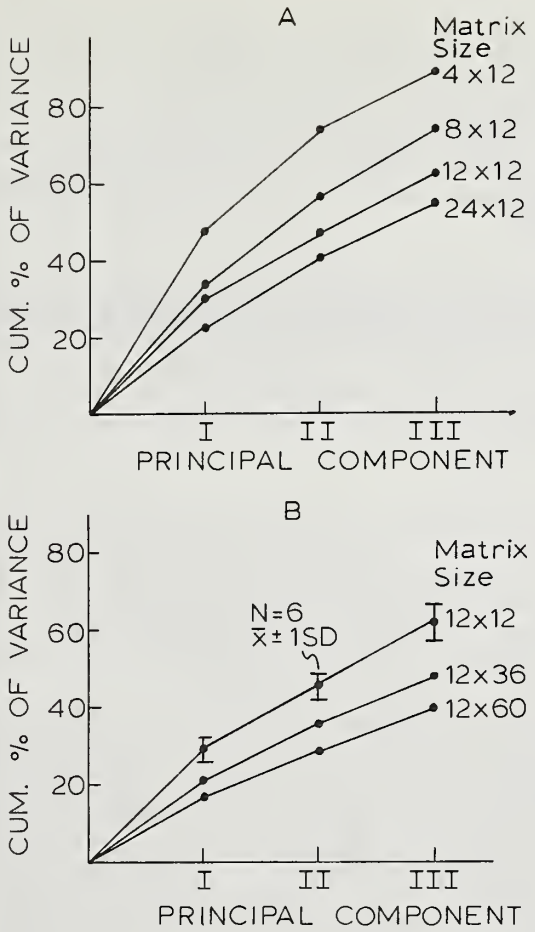


Figure 3. Cumulative percent of variation accounted for in principle component analysis of random number tables of various sizes.

accounted for by the second (real - 42%; random - 50%) and third (real - 34%; random - 64%) components is greater for the random data. While the pattern is of interest, we know of no statistical procedures that can be used to test its significance. At the least, the pattern emphasizes the need for caution when using PCA on small matrices.

Even if the second and third components explain a greater proportion of the remaining variation in real data than in random numbers, it may be difficult to apply biological interpretations to these axes if they are equal. The sphericity test indicates that the second and third components are not significantly ($P > 0.05$) different in any of the studies in table 2. One is tempted to conclude that, in bird/habitat data, the first axis reflects the presence of a primary ellipsoid while subsequent axes are nearly as spherical as a random numbers matrix of the same dimensions. If this is true, interpretations of components subsequent to the first may be invalid; at the least, they should be tested by the sphericity test before they are interpreted in great detail.

THE EFFECTS OF MATRIX SIZE

Obviously, the amount of variation explained from random numbers varies with number of derived axes since 100% of the variation must be extracted from a number of axes equal to or less than the number of variables in the analysis. For example, a matrix of 10 vegetation variables with 24 bird species must yield 100% variation in 10 or fewer axes. Typically, all variation is extracted with fewer components than variables. When there are few variables, the problem becomes especially critical (e.g., 89% on the first three principal axes in a 4 x 12 matrix; fig. 3a).

Similarly, when the number of variables is held constant but the number of cases varies, the percent of variance accounted for declines with increasing matrix size (fig. 3b). Note also that variability among random number matrices of the same size is rather small (fig. 3b).

DISCUSSION

We must emphasize that our purpose is not to convince the reader to avoid using principal components analysis. Rather, we hope to inspire more critical evaluation of the uses and misuses of PCA. When the amount of variation accounted for from an axis or set of axes is similar to the amount from random numbers, we question the biological validity of interpreting those axes.

Canonical correlation (CanCorr) analysis is similar to PCA in its analytical procedures except that CanCorr is developed to analyze the relationship between two data sets. In addition, CanCorr includes a test of significance of the canonical variates. We have applied the test to random number matrices of a variety of sizes; no significant ($P > 0.05$) canonical variates were found. Most biologists would probably not try to apply biological interpretations to such results. Use of a similar decision-making process is needed in the use of PCA.

Finally, we note that similar results from principal components analysis of random and real data do not necessarily mean that biologically important relationships are not present. However, it is important to use sound biological judgment in the interpretation of the results. The final test of the procedure in any circumstance is the biological insight developed. That is, the results of PCA should not be used simply for post facto interpretations, but rather to aid researchers in generating reliable predictions and viable hypotheses which are testable with additional research.

ACKNOWLEDGMENTS

J. Blake and I. Schlosser made helpful comments on an early draft of the manuscript. M. Tatsuoka helped to clarify some of the problems associated with the use of the sphericity test.

LITERATURE CITED

Pimentel, R.A. 1979. Morphometrics. The multivariate analysis of biological data. 276 p. Kendall/Hunt Publishing Co., Dubuque, Ia.

DISCUSSION

JAKE RICE: In defense of the use of these methods, in this case PCA, I can at least say that the field of statistical ecology has progressed to the point where one usually cannot get a paper published just by doing a PCA on a data set and writing up the result. If the factor loadings make biological sense, like Smith's forest moisture gradient (Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark Watershed. Ecology 58: 810-819), then one has some results to work with. If the loadings and the scores do not make sense, one goes back, or should go back, and look hard at the system; obviously one's understanding (and data base) of the system are incomplete.

If you also randomly assigned habitat names to "variables" and site names to "samples," in your random matrix; factored it; and could put a plausible biological explanation on the results, then I would be surprised.

JIM KARR: Many users of PCA and other multivariate procedures have progressed beyond the "fishing expedition" approach but others have not. The questions posed in this paper are directed toward both groups--primarily as a caution about interpretation and use of results. They should not be construed as suggestions to avoid the procedures and the new insights they can produce if used wisely. A problem inherent in PCA is the lack of objective statistical tests to determine the significance of the patterns which "make sense." Note that both statistical and biological significance are of concern; they may not be the same.

We clearly disagree with your final point. I am confident that most biologists could generate plausible "post-facto" explanations for high loadings after randomly assigning habitat names to the "variables" in random number tables. The small sample of examples that I have seen clearly demonstrate the ingenuity of biologists in that regard.

CHARLES SMITH: We typically assume that there is a pattern of some kind; and when none is detected, we question the validity of the techniques. What if the most easily detected pattern is one in time, not space, and the spatial case is not detectably different from a random pattern, in the absence of a time-series sample?

JIM KARR: This is a real problem. Indeed, I have fallen into that trap myself. After one year of study in forests in Panama, I concluded that individuals and species moved extensively among microhabitats in lowland forest. Long-term data show that to be true but the subtlety of habitat selection is exceptional. The use of specific microhabitats varies between seasons and years in ways that are quite predictable from knowledge of environmental conditions (wet vs. dry dry season) at the time. Lumping of data for many kinds of statistical analysis can result in the error you mention (i.e., assuming a lack of pattern when a pattern, in fact, exists); that simply reinforces the point that statistical analyses should always be guided by biological knowledge (insight).

Special Session: Sampling Avian Habitats

RATIONALE AND TECHNIQUES FOR SAMPLING AVIAN HABITATS: INTRODUCTION[1]

James R. Karr[2]

Three more or less distinct (but overlapping) stages might be recognized in the development of studies of avian habitats. The first "Catalog Stage" began with efforts to identify birds and determine their phylogenetic and biogeographic affinities. Habitat descriptions were generally cursory and nonquantitative. For example, the monumental Manual of Neotropical Birds (Blake 1977) describes the habitat of the red-tailed hawk (Buteo jamaicensis) as "Mainly woodland and semiopen country."

During the second "Natural History Stage" scientists were interested in general biology or natural history of species: nest type and location, food habits, clutch size, incubation period and general habitat were of primary interest. Few efforts were made to provide quantitative information on avian habitats; the main focus was the bird itself. The classic works of Nice (1937) on the song sparrow and Skutch (e.g. 1969) on neotropical birds are examples which focus on the natural history of birds, including nonquantitative studies of their habitats.

In the third "Ecology of Habitat" stage emphasis shifted to interest in both the birds (often as communities) and their habitat (vegetation type). The importance of foliage structure in determining avian use of habitat has long been recognized (Merriam 1890, Grinnel 1917, Lack 1937, Pitelka 1941, Kendeigh 1945, Svardson 1949). The first quantitative and graphical demonstration that vegetation of increasing height and complexity typically supports increasingly diverse avifaunas was provided by MacArthur and MacArthur (1961). Many refinements of this approach have been developed in the past two decades. Some are appropriate for consideration of single species, while others are more appropriate for community level studies.

Increased emphasis on quantitative approaches to the study of avian habitats has precipitated some innovative uses of quantitative methods. Use of information theory was pioneered by Robert MacArthur and has continued with the work of others (Recher 1969, Karr and Roth 1971, Blondel et al. 1973). Other efforts to demonstrate a relationship between foliage height diversity and bird species diversity have been less successful. Some failed because of inappropriate measures of habitat structure. Others failed in habitats where the rules of plant geometry and distribution are different. For example, foliage volume is important in some situations (Sturman 1968, Laudenslayer and Balda 1976), while life form diversity of plants is more important in others (Tomoff 1974). Further, precision of relationships between habitat structure and avian diversity is often less precise when narrow ranges of habitat structure are examined (Lovejoy 1974, Willson 1974).

An array of multivariate statistical procedures also are being used in studies of bird-habitat relations. In general, these techniques (principal components, discriminant function analysis, reciprocal averaging, canonical correlation) have the attribute of reducing many variables to a small set of complex variables. Other attractive features include easy development of graphical presentations and studies of how variation in one variable affects birds when other variables are held constant.

Multivariate methods are not without problems, however. First, like most univariate methods, they are merely descriptive procedures. They are essentially correlation techniques and cannot be used to determine casual (ultimate or even proximate) relationships between birds and their habitats. Although complex data transformations result from use of multivariate procedures, it is not always clear how to extract biological (vs. statistical) meaning from these correlations. Further, the caution that ecology is the art of describing the obvious is also

---

[1] Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2] Professor, Department of Ecology, Ethology, and Evolution, University of Illinois, 606 E. Healey, Champaign, IL 61820.

relevant. There is a tendency to use multivariate methods to demonstrate phenomena which are already obvious from univariate studies. The true test of a procedure is its ability to generate new predictions about biological phenomena, to go beyond the obvious. Application of multivariate methods cannot substitute for in-depth consideration of the biology of the organisms under investigation. Careless use of these procedures has stimulated one biologist to comment that so much mathematical formality combined with so much ecological casualness is puzzling (Beals 1972).

THE DEFINITION OF HABITAT

Use of the concept of habitat by a single investigator (and among investigators) is often inconsistent (Karr 1980). At least three major meanings of habitat commonly can be discerned:

1. Habitat = vegetation type (grassland, forest, etc.).
2. Habitat = the living and non-living surroundings of an organism.
3. Habitat = specific horizontal (vegetation configuration) or vertical (twig angle, leaf density, etc.) components of vegetation structure. These are often referred to as the microhabitat of the species.

Obviously, it is not possible to address all of these and other difficulties involved in the analysis of avian habitats in a few short papers. However, the papers that follow outline many of the problems and propose some innovative solutions as guides for the future. Presentations are organized around three major questions:

1. Why do we measure habitat?
2. What habitat variables should be measured?
3. How do we measure those variables?

The "why" question must generally be answered first. The answer will place constraints on the "what" question (and so on to the "how" question).

THE "WHY" QUESTION

In addressing this question Rotenberry provides the simple answer: It works. He argues that it works because habitat forms the background upon which all adaptive patterns are expressed. Further, it works for a wide range of studies. Successes using the habitat approach include both basic and applied objectives; single species, several species, or entire communities. The need for quantitative habitat assessment is especially critical in efforts to preserve endangered species. Since the type of "why" question helps define approach, sampling protocols, and nature of data analysis in any study, dealing with the "why" question is an important first step.

THE "WHAT" QUESTION

Since not all relevant parameters can be measured in any study, researchers must restrict the "what" to strike a balance between time and money constraints and the need of a sound research design with high probability of developing useful conclusions. Exact variables to be measured, as well as how precisely they will be measured, vary with study objectives. The emphases of the two contributions here are toward theoretical and applied considerations. In the applied situation, a reliable correlation between a habitat variable and avian patterns of interest may often be sufficient for management decisions. Whitmore's presentation makes the point that sound research design directed toward a distinct management objective can yield results of major significance for management to enhance wildlife populations.

In a theoretical context, it may be more important to understand the causal links relating habitat and avian species or communities. Thus, a theoretician may need to sample insect (or other food) abundances in space and time to understand complex relations between physical environment, vegetation structure, food supply, and avian habitat use. This more complex picture may be necessary for a rewarding exploration of ecological and evolutionary causes of patterns in avian species and communities. In addressing this question, Holmes outlines examples to show the need for more rigorous procedures in determining which habitat variables should be measured. He argues strongly for the need to consider natural history of the study organism(s) in determining variables to be measured. Based on this principle, he outlines several difficulties with current approaches and cautions that more detailed and precise information is needed before truly scientific management can be accomplished.

THE "HOW" QUESTION

The answers in a specific case to the "why" and "what" questions narrow the options in the "how" question. In some cases, superficial (perhaps even nonquantitative) measures of very general factors are all that is required; in other cases, detailed and comprehensive measurements may be necessary. Both biological and statistical constraints must be kept in mind at this stage. Noon's extensive experience with sampling avian habitats leads him to propose a variety of specific protocols for use in several circumstances. His proposals are a valuable mix of techniques which have proved successful for a number of researchers. He calls for efforts to standardize some procedures to facilitate comparisons among studies. While there is some merit to this point, I urge caution in too firmly establishing a sampling protocol. It is premature for protocols to be "chipped-in-stone" in such a rapidly developing area. Some adoption of sampling conventions should be attempted, but the way for improvement of those procedures should be kept open as knowledge of the relations between

birds and their habitats is improved. In the final paper of the series, Johnson urges biologists to make study design decisions with a clear view of the objectives and goals of each specific study. He contrasts exploratory and confirmatory research, while urging that conclusions be based on sound scientific design (hypothesis testing, etc.) and not fishing expeditions for significant correlations.

Finally, I would like to reiterate the point that all insights will not depend on use of complex multivariate procedures. Although that is the general subject of these proceedings, it is important that we reflect on the reality that the final measure of value for a study or technique is whether or not it yields insight into the dynamics of the ecological systems that we are studying.

LITERATURE CITED

Beals, E.W. 1972. Ordination: mathematical elegance and ecological naivete. Journal of Ecology 60:23-35.

Blake, E.R. 1977. Manual of neotropical birds. volume 1. 674 p. University of Chicago Press, Chicago, Ill.

Blondel, J., C. Ferry, and B. Frochot. 1973. Avifaune et vegetation. Essai d'analyse de la diversite. Aluda 41:63-84.

Grinnell, J. 1917. Field tests of theories concerning distributional control. American Naturalist 51:115-128.

Karr, J.R. 1980. Strip-mine reclamation and bird habitats. p. 88-96. In DeGraaf R.M., and N.G. Tilghman, compilers. Management of western forests and grasslands for nongame birds: Proceedings of a workshop [Salt Lake City, Utah, Feb. 11-14,1980]. USDA Forest Service General Technical Report. INT-86, 535 p. Intermountain Forest and Range Experiment Station, Ogden, Ut.

Karr, J.R., and R.R. Roth. 1971. Vegetation structure and avian diversity in several New World areas. American Naturalist 105: 423-435.

Kendeigh, S.C. 1945. Community selection by birds in the Helderberg Plateau of New York. Auk 62:418-436.

Lack, D. 1937. The psychological factor in bird distribution. British Birds 31:130-136.

Laudenslayer, W.P., and R.P. Balda. 1976. Breeding bird use of a pinyon-juniper-ponderosa pine ecotone. Auk 93:571-586.

Lovejoy, T.E. 1974. Bird diversity and abundance in Amazon forest communities. Living Bird 13:127-191.

MacArthur, R.H., and J.W. MacArthur. 1961. On bird species diversity. Ecology 42:594-598.

Merriam, C.H. 1890. Results of a biological survey of the San Francisco Mountain region and desert of the Little Colorado in Arizona. U.S. Department of Agriculture, North American Fauna 3:1-136.

Nice, M.M. 1937. Studies in the life history of the song sparrow. I. Transactions of the Linnean Society of New York 4:1-247.

Pitelka, F.A. 1941. Distribution of birds in relation to major biotic communities. American Midland Naturalist 25:113-137.

Recher, H.F. 1969. Bird species diversity and habitat diversity in Australia and North America. American Naturalist 103:75-80.

Skutch, A.F. 1969. Life histories of Central American birds. III. Pacific Coast Avifauna 35:1-580.

Sturman, W.A. 1968. Description and analysis of breeding habits of the chickadees, Parus atricapillus and P. rufescens. Ecology 49:418-431.

Svardson, G. 1949. Competition and habitat selection in birds. Oikos 1:157-174.

Tomoff, C.S. 1974. Avian species diversity in desert scrub. Ecology 55:396-403.

Willson, M.F. 1974. Avian community organization and habitat structure. Ecology 55:1017-1029.

WHY MEASURE BIRD HABITAT?[1]

John T. Rotenberry[2]

Abstract.--The rationale behind studies that attempt to assess bird/habitat relationships quantitatively is structured around the premise that a bird species' habitat selection is of adaptive significance. This rationale was originally posited by Grinnell and led to the development of the concept of "niche". Although the definition and use of the term "niche" have undergone substantial expansion, it seems clear that habitat variables (especially those associated with vegetation structure) do represent an integral part of bird species' niches. By comprising either proximate or ultimate factors to which species must respond, habitat forms the background upon which all adaptive patterns are expressed. Thus a full understanding of the evolution of species' ecological attributes can come only in association with precise quantitative descriptions of environmental conditions.

Key words: Adaptation; birds; Grinnell; habitat measurement; habitat selection; niche.

INTRODUCTION

With maturation of natural history as a science, it is apparent that information of a general descriptive or quantitative nature about habitat is absolutely essential for a full understanding of patterns of life history, adaptation, and evolution of any species. This seems especially apparent for birds. As I perceive the role of this paper, it is not to state that the reason you should measure bird habitat is because it "works," then proceed to provide you with an exhaustive list of examples; indeed, many of the contributors to these proceedings are responsible for those examples, and their reiteration would prove superfluous. Instead, I shall attempt to summarize briefly some of the rationale behind why we think it "works," structured around the notion that an organism's habitat reflects aspects of its adaptations.

NICHE CONCEPTS

With birds, perhaps more than for any other taxa, there has been a historical as well as a theoretical component to why we gather habitat data. The seminal paper in bird/habitat relationships, and indeed for much of ecology today, is Joseph Grinnell's classic study of the California thrasher (Toxostoma redivivum) (Grinnell 1917). Grinnell introduced two important concepts that have formed the foci for the way we think about birds and their habitat relationships. The first of these, of course, was the creation of the term "niche" and its conceptualization as a close association between distributional patterns of a species and the underlying environmental conditions. The second was the notion that niche relationships were important not only in telling us about adaptation

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Research Associate, Shrubsteppe Habitat Investigation Team, Department of Biology, University of New Mexico, Albuquerque, NM 87131. Current address: Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403.

and natural history of whatever organism we are describing, but also in revealing aspects of its relationships to other organisms and, ultimately, of the structure of the community in which it resides. The latter idea has received more emphasis in a theoretical context, largely under the construct of the n-dimensional hypervolume model of Hutchinson (1958). Both of these concepts, the niche as habitat and as something that reflects individual adaptation and community organization, form the basis for the sorts of questions addressed in these proceedings.

Grinnell's original concern was to explain the rather restricted ecological distribution of the California thrasher. He thought that reasons could be found in the various physiological and behavioral adaptations of the bird to a narrow range of environmental conditions. It seemed evident that the nature of these critical conditions was to be learned through an examination of the bird's habitat. From this Grinnell went on to relate various aspects of thrasher biology and distribution to certain physiognomic, floristic, faunistic, and climatic features of the environment, and described these as the "niche-relationships" of the species.

Since Grinnell, niche has been redefined several times, with most of these reformulations stressing some aspect of the functional role of the organism within the community (e.g. Elton 1927). Some have even argued for a strict separation of a species' habitat and function, with the former representing environmental relations and the latter the "true" niche (Whittaker et al. 1973). It seems more probable, however, that these ideas simply represent ends of a continuum, with substantial intergradation between the two. Insofar as birds can be shown either to partition or to select different subsets of habitat within a community (e.g. Cody and Walter 1976), those differences in habitat among them will provide indications of the differences in their functional roles. It seems clear, then, that habitat variables do represent an integral part of a bird species' niche regardless of the presumed rigor of one's definition of the niche in general; indeed, numerous aspects of niche theory have been elaborated in just this context (Shugart and Patten 1972).

HABITAT SELECTION

Of course, the ornithologist's approach to the concept of habitat as niche has evolved through the years as well, but it still embodies the basic assumption that a predictable relationship exists between the occurrence of a species and certain characteristic habitat requirements. The current approach is often to focus on the floristic or vegetational structural aspects of avian habitat, what James (1971) has termed the "niche-gestalt." The concept of niche-gestalt is predicated on the theory that there is some basic configuration or pattern in the environment that an individual animal will seek out and settle in. This habitat selection process may be based on a specific search image, early learned experience, the particular genetic make-up of the individual, or any combination of these factors (Klopfer 1970). Presumably, then, habitat selection is an evolutionarily derived mechanism that insures that individuals seek out and remain in the particular environment to which they are adapted. This implies, in turn, that the result of habitat selection, that is, the association of a particular species with a particular subset of an otherwise vast environmental landscape, will reflect major aspects of a species' ecology and behavior.

The recognition stimuli that induce a bird or any other animal to select a particular habitat may often appear to be unrelated to the actual survival and successful reproduction of that organism (Hilden 1965). Such proximate factors achieve importance less through direct influences on the animal's biology than through correlation or association with the ultimate factors that are influential, in an adaptive sense. It is the latter that are essential to survival and constitute the underlying selective pressures to which the species is adapted. In many respects, then, the niche-gestalt is that set of proximate factors that elicits a habitat selection response (Smith 1977). These factors are presumed, in turn, to provide the birds with predictive evidence of ultimate factors. For birds, the physical structure of the habitat has long been considered to be an important proximate niche dimension, either directly, as it provides shelter, nesting substrate, or protection from predators, or indirectly, as it provides cues to the potential availability and diversity of prey (Wiens 1969).

In dealing with habitat physiognomy, ornithologists generally consider the measured parameters to be representative of proximate aspects of the niche. From the readily observable proximate relationships, however, we then inductively create hypotheses or models that often deal with functional elements of the niche, or the species' role within the community. Again the "habitat" and "functional role" aspects of the niche concept seem intertwined rather than exclusive alternatives. If there is a relatively good correspondence between the proximate and ultimate factors, then from these hypotheses and models we can deductively create testable predictions about other relationships between birds and their environment, or even between sets of bird species.

Although the very concept of "niche" has been challenged by some ecologists as being trivial (e.g., Ricklefs 1973: 522), it is clearly on the merit of its usefulness as a model or predictive tool that it should stand or fall. It seems quite clear for birds, at least, that the notion of habitat as niche has been extremely useful. Perhaps the most direct manifestation of this is the relative ease with which numbers of individuals or relative densities of many bird

species may be predicted from measurement of habitat variables (e.g., Anderson and Shugart 1974, Robbins 1978). Likewise, the simple presence or absence of a species may often be strongly correlated with a suite of environmental measures (e.g., Smith 1977). Such relationships have been empirically and statistically verified many times, and some bird/habitat relationships are so strong that field identification guides may use habitat occupancy as a key character (Robbins et al. 1966). If in fact habitat selection does represent a major evolutionary component of avian natural history, then these correlations may allow us to make predictions about the adaptations of particular species. We may also identify species that depend on rather specific habitat conditions (what we call narrow-niched species, or specialists) versus those that are associated with a wide variety of conditions (broad-niched, or generalist species). Through this relationship we may again infer adaptive relationships.

Such correlations and associations between single species of birds and habitat variables are also of substantial practical value. They allow us to predict with varying degrees of accuracy the response of a species to natural or artificial habitat alterations. They permit us to identify species that, because they are relatively specific in their habitat requirements, are most likely to undergo major population declines if certain portions of that habitat are altered. And finally, they allow us to identify environmental management practices that may increase the amount of appropriate habitat for species that have become endangered or rare as a result of previous habitat loss.

COMMUNITY COMPOSITION

On a large scale, niche theory and the distribution of species along environmental gradients allow us to make predictions about the coexistence or co-occurrence of species in communities. To the extent that species are distributed along a habitat gradient more or less independently of one another, the community composition of a site is predicted merely by its location along that gradient (Rotenberry and Wiens 1980). Niche theory suggests, however, that under certain environmental conditions species may not vary independently from one another, but instead that inter-populational interactions, such as competition, will contribute to the nonrandom distribution of species along habitat gradients (Terborgh 1971). This, of course, has proven to be a fertile area for theoretical speculation and prediction, and the role of competition in organizing bird communities, expressed through either habitat displacement or, conversely, coexistance mechanisms, has been investigated extensively (see especially Cody 1974). Regardless of whether one chooses to look at community composition as a result of the independent distribution of species on environmental continua or the product of intense competitive interactions leading to tightly ordered distributional patterns, it seems clear that the nature of the underlying habitat will have a profound effect on community composition. This observation is perhaps best exemplified by what has become virtually a truism in avian ecology: bird species diversity may be regularly predicted from vegetational complexity, as expressed through either plant species diversity (Lovejoy 1974), life form diversity (Tomoff 1974), or structural diversity (MacArthur and MacArthur 1961). Of course, it will be only through careful measurement of appropriate habitat variables coupled with knowledge of species' biologies that we will be able to distinguish the independent versus competitive alternatives.

CONCLUSION

Quite beyond all these other reasons that I have outlined as providing a rationale for measuring bird habitat is the simple observation, as I indicated at the outset, that the measurement of habitat does seem to "work" for birds. By "work," I mean that there appear to be regular, repeatable patterns of associations or correlations between birds and habitat variables, regardless of any theoretical expectations or interpretations. Empirically this is verified by the literally hundreds of papers that have appeared since Grinnell that demonstrate nonrandom occupancy of habitat. These documentations may range from things as simple as the correlation between sagebrush coverage and sage sparrow (Amphispiza belli) density in the Great Basin (Rotenberry and Wiens 1978), on up through a detailed elaboration of the relationship between foliage height diversity and bird species diversity in a tropical forest (Karr and Roth 1971).

The message I wish to convey is simple: if we are to discuss any bird species' ecology in an adaptive context, information about its habitat is essential. This is because habitat forms the background on which all adaptive patterns are expressed. Virtually all attributes of a species, from its internal physiology on up through its interaction with other members of its community, have evolved for certain environmental conditions. Without knowledge of those conditions, which is expressed through our quantitative or qualitative description of habitat, the adaptive nature of these attributes is unknown. It seems apparent, therefore, that the necessity of defining these environmental conditions will result in the continued intertwining of bird populations and habitat measurements throughout all phases of avian ecology.

ACKNOWLEDGMENTS

John Wiens, Linda Heald, and Dennis Heinemann provided constructive criticism of an earlier version of this paper. Recent support has been provided by NSF 75-11898 to John Wiens.

LITERATURE CITED

Anderson, S.H., and H.H. Shugart. 1974. Habitat selection of breeding birds in an east Tennessee deciduous forest. Ecology 55:828-837.

Cody, M.L. 1974. Competition and the structure of bird communities. 318 p. Princeton University Press, Princeton, N.J.

Cody, M.L., and H. Walter. 1976. Habitat selection and interspecific interactions among Mediterranean sylviid warblers. Oikos 27:210-238.

Elton, C. 1927. Animal ecology. 209 p. Sidgwick and Jackson, London.

Grinnell, J. 1917. The niche-relationships of the California thrasher. Auk 34:427-433.

Hilden, O. 1965. Habitat selection in birds. Annales Zoologici Fennici 2:53-75.

Hutchinson, G.E. 1958. Concluding remarks. Cold Spring Harbor Symposium on Quantitative Biology 22:415-427.

James, F.C. 1971. Ordination of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Karr, J.R., and R.R. Roth. 1971. Vegetation structure and avian diversity in several New World areas. American Naturalist 105: 423-435.

Klopfer, P. 1970. Behavioral ecology. 173 p. Duke University Press, Durham, N.C.

Lovejoy, T.E. 1974. Bird diversity and abundance in Amazon forest communities. Living Bird 13:127-191.

MacArthur, R.H., and J.W. MacArthur. 1961. On bird species diversity. Ecology 42:594-598.

Ricklefs, R.E. 1973. Ecology. 861 p. Chiron Press, Newton, Mass.

Robbins, C.S. 1978. Determining habitat requirements of nongame species. Transactions North American Wildlife and Natural Resources Conference 43:57-68.

Robbins, C.S., B. Bruun, H.S. Zim, and A. Singer. 1966. Birds of North America. 340 p. Golden Press, New York, N.Y.

Rotenberry, J.T., and J.A. Wiens. 1978. Nongame bird communities in northwestern rangelands. O. 32-46. In DeGraaf, R.M., technical coordinator. Nongame bird habitat management in the coniferous forests of the western United States: Proceedings of a workshop [Portland, Oregon, Feb. 7-9, 1977]. USDA Forest Service General Technical Report PNW-64, 100 p. Pacific Northwest Forest Experiment Station, Portland, Ore.

Rotenberry, J.T., and J.A. Wiens. 1980. Habitat structure, patchiness, and avian communities in North American steppe vegetation: a multivariate analysis. Ecology 61:1228-1250.

Shugart, H.H., and B.C. Patten. 1972. Niche quantification and the concept of niche pattern. p.284-327. In Patten, B.C., editor. Systems analysis and simulation in ecology. Volume II. Academic Press, New York, N.Y.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58:810-819.

Terborgh, J. 1971. Distribution on environmental gradients: theory and preliminary interpretation of distributional patterns in the avifauna of the Cordillera Vilcabamba, Peru. Ecology 52:23-40.

Tomoff, C.W. 1974. Avian species diversity in desert scrub. Ecology 55:396-403.

Whittaker, R.H., S.A. Levin, and R.B. Root. 1973. Niche, habitat, and ecotope. American Naturalist 107:321-338.

Wiens, J.A. 1969. An approach to the study of ecological relationships among grassland birds. 93 p. Ornithological Monographs.

THEORETICAL ASPECTS OF HABITAT USE BY BIRDS[1]

Richard T. Holmes[2]

Abstract.--For the theoretical ecologist interested in understanding why correlations exist between the occurrence of bird species and certain habitat variables, it is necessary to know how birds use their habitats. Such information also provides the basis for choosing appropriate habitat variables to measure. These points are illustrated by observations of birds in a deciduous forest in New Hampshire, in which different species of trees are found to be important habitat parameters.

Key words: Bird foraging; factor analysis; Hubbard Brook; vegetation structure.

INTRODUCTION

A major goal of the field of avian ecology is to develop an understanding of the factors that determine the patterns of occurrence, distribution, and abundance of birds. With respect to habitat, the avian ecologist is specifically concerned with why birds occur where they do, why there are correlations between bird species distributions and certain habitat variables, and particularly what are the ecological and evolutionary causes of these relationships and patterns. Since it is important to understand the causal relationship underlying the observed patterns before devising or implementing a management plan, such information is also of direct use to the wildlife ecologist/manager who is interested in manipulating bird habitats.

In this paper, I review briefly some of the protocols and procedures used in studies of avian habitats in terrestrial, mainly forested environments, and recommend that more attention be focused on the behavior of the animals themselves. By knowing more about how birds use their habitats, the ecologist can make better decisions about which habitat variables to measure.

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Professor, Department of Biological Sciences, Dartmouth College, Hanover NH 03755.

ASSESSING BIRD HABITATS: A SYNOPSIS AND CRITIQUE

The usual way to examine the relationship between birds and their habitats has been to select a set of habitat characteristics and to relate these, frequently with the use of multivariate techniques, to the presence of bird species in those habitats. The habitat variables to be measured are often chosen rather arbitrarily, albeit based on the investigator's biological intuition as to what might or should be important to the birds.

The variables chosen often include some measure of plant species composition (usually the number of tree species) and several features that describe vegetation structure (often deciduous vs. coniferous foliage, tree size classes, canopy height, percent canopy cover, percent ground cover, etc.). Such measurements are either made 1) systematically through a study plot and then related to the avifaunal composition of that area (e.g. MacArthur and MacArthur 1961, James and Shugart 1970, Willson 1974, Titterington et al. 1979), or 2) on the territories of individual birds (e.g. James 1971, Whitmore 1975, 1977, Smith 1977). For the latter, the standard procedure has been to choose a song perch or nest site as the center of a 0.04 ha circle (11.3 m in radius) in which the habitat is then measured. One such circle is taken per territory, and a number of territories of each species are measured. The data so obtained are considered to characterize the places where the birds live, and have led to biologically meaningful correlations or

ordinations (e.g. James 1971, Whitmore 1975, 1977, Smith 1977). Furthermore, with appropriate statistical techniques (see other papers in these proceedings) the habitat variables that are best correlated with the presence of a particular species can be identified.

Putting aside for a moment the problems of arbitrarily chosen variables, there are a number of difficulties with these sampling techniques. For instance, since an area of 0.04 ha represents only a relatively small portion of the territories of most passerine birds, can habitat measurements made in a 22 m diameter circle around a single song` perch adequately represent the habitat occupied by that bird? I contend, for reasons given below, that they may not. Also, how many territories must be measured to provide a statistically valid sample for representing the habitat occupied by the species? Most studies so far have taken measurements from an average of 10 to 20 territories per species, but none has yet evaluated how sufficient these sample sizes are.

Finally and more pertinent to the subject of this paper, these techniques provide little information, if any, on the reasons why the measured habitat variables might be important to the birds. Nor do they provide a basis for evaluating whether the most appropriate variables have been chosen in the first place. Thus, not only do the existing sampling methods need further development and testing, but the basic rationale for deciding what habitat components to measure in the first place requires re-evaluation.

HOW DO BIRDS USE THEIR HABITATS?

What then should we measure? What habitat variables are important to birds? To answer such questions, it is necessary to understand how birds utilize their habitats and what habitat components ultimately influence the survivorship and reproductive success of these birds. Although such information must eventually come from intensive population studies, considerable insight can be obtained from detailed observations of birds in their habitats.

By occupying a particular habitat, birds gain more than just a place to live (Hilden 1965). They obtain places to hide from predators, to escape vagaries of weather, to display, roost, nest and forage. The relative importance of these various functions differs from species to species and habitat to habitat. For example, suitable nest sites may be particularly important for desert birds (Tomoff 1974) or for cavity-nesters whose presence, abundance and reproductive success in a habitat will be determined in large part by the availability of suitable nesting sites (Conner 1978).

Perhaps the way in which habitats are used most intensively and extensively by birds, especially those inhabiting forests, is for foraging. With a relatively large proportion of their daily activities spent searching for food, birds move frequently from place to place and constantly scan plant surfaces, including tops and bottoms of leaves, twigs, branches and tree boles, for prey. Many of these birds move long distances during each foraging bout and at least in temperate forests, range widely from the forest floor to the upper parts of the canopy.[3]

To illustrate how observations on bird foraging behavior can provide information about what habitat variables might be important, I summarize here briefly some results from a study of birds in the Hubbard Brook Experimental Forest, West Thornton, N.H. The methods and analytical procedures have been described in detail by Holmes et al. (1979a). Basically, data were collected on the microhabitat use by birds foraging in this northern hardwoods forest during the breeding season, late May through mid July 1974-1976. Information was recorded on the height of foraging, the tree species on which foraging occurred, substrates to which foraging maneuvers were directed, and the kinds of foraging maneuvers employed. For purposes here, only the 11 bird species that forage primarily among the foliage of the forest canopy and the 20 appropriate foraging-related characters (table 1; Holmes et al. 1979a) are considered. The 20 x 11 "species" matrix (Q-technique, Sneath and Sokal 1973) was used to calculate the Euclidean distances between all combinations of the 11 species in the multi-dimensional space defined by the 20 foraging characters. This distance matrix was then subjected to hierarchical cluster analysis for purposes of illustrating species relationships (fig. 1). The transposed 11 x 20 'character' matrix (Q-technique, Sneath and Sokal 1973) was used for a varimax rotated factor analysis, as described by Holmes et al. (1979a).

The dendrogram in figure 1 groups the 11 bird species on the basis of their similarities or differences in microhabitat use/foraging behavior. There are clearly two major groups of foliage-foraging species in this community, each with two to several subgroupings. The relative importance of the characters that account for this pattern can be determined from the factor analysis, which weights the variables by their relative contributions to the total community pattern and reduces a large number of variables to a smaller number of identifiable factors (`Cooley and Lohnes 1971, Bhattacharyya 1981). In this case, the first four factors account for 79% of the community variance (table 1), and illustrate how the birds differentially utilize foraging location, substrate and tree species. Thus, factor I largely accounts for the first major subdivision in the dendrogram. It separates those species that primarily hover for prey on leaves, hawk insects from the air and associate their

[3]Holmes, R.T., unpublished data on file, Dartmouth College.

Table 1. The rotated factor pattern showing the most heavily weighted factors, either positive or negative, for each of 20 foraging characters (see Holmes et al. 1979a for further details).

Factors	I	II	III	IV
Eigen roots	7.24	4.27	2.47	1.71
Factor contribution to community variance (%)	36.22	21.36	12.39	8.56
Cumulative %	36.22	57.58	69.97	78.53
1 Hover leaf	0.889			
2 Glean leaf		0.728		
3 Hover branch				
4 Glean branch	-0.673			
5 Hover twig				-0.951
6 Glean twig	-0.929			
7 Hawk (air)	0.589			
8 Hover trunk		0.806		
9 Glean trunk	-0.805			
10 Proximal		0.946		
11 Distal		-0.971		
12 Beech	0.792			
13 Maple	0.743			
14 Birch	-0.920			
15 Ash	-0.499			
16 Shrub	0.799			
17 Conifer			-0.809	
18 Height (x)		-0.616		
19 Height SD			0.893	
20 Body size				

foraging with sugar maple (Acer saccharum) and beech (Fagus grandifolia) from those that glean prey, largely from twigs, branches and boles of trees and often forage on yellow birch (Betula allegheniensis) and white ash (Fraxinus americanus). The other factors illustrate further how species differently utilize the foraging environment by segregating those that glean prey from leaves, hover at tree boles and forage primarily in the inner (proximal) portions of the trees from those that forage distally along the branches, usually high in the canopy.

The finding that these birds are differentially using tree species and foraging substrates suggests that they are responding to, or are influenced in some way by, these habitat components. Indeed, it is very likely that these elements are closely linked, in that different kinds of foraging behavior may be required to exploit prey occurring on particular types of substrates which may vary in color, texture or frequency on different species of trees. This is supported by a more detailed study of bird foraging behavior which indicates that these insectivorous birds in the northern hardwood forests at Hubbard Brook have distinct preferences and/or aversions for certain species of trees while foraging (Holmes and Robinson 1981). This is attributed at least in part to differences among the tree species in arthropod abundances and in foliage arrangements that influence how birds search for and capture prey from plant surfaces. To determine the former, we utilized detailed information on insect abundances obtained from an intensive sampling of bird food resources (Holmes and Robinson 1981, Holmes et al. 1979b). For the latter, we found that bird species which typically glean arthropods have the strongest tree species preferences, in this case for yellow birch and for conifers (Holmes and Robinson 1981). On these plants, the leaves are arranged close to the branches so that a standing bird can reach prey situated on leaf surfaces. In contrast, in maple or beech, leaves are arranged either on flat sprays or at the ends of long petioles, and birds must fly and snatch their prey or hover for them at leaf surfaces. In addition, there is

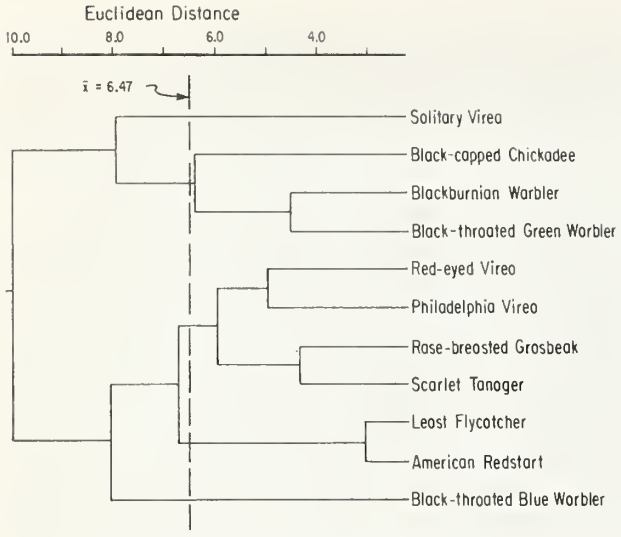


Figure 1. Dendrogram indicating the similarities and differences in microhabitat use among the 11 bird species at Hubbard Brook that forage for arthropod prey among foliage (for detail of technique, see text and Holmes et al. 1979a).

preliminary evidence that abundance and distribution of some of these bird species may be linked to particular tree species or at least to trees with similar physiognomic characteristics (Holmes and Robinson 1981).

The important point here is that observation of foraging birds has led to the realization that different species of broad-leaved trees in a primarily deciduous forest can be important habitat components. Although the potential importance of deciduous vs. coniferous foliage has been previously recognized (Balda 1969, Franzreb 1978) and occasionally incorporated into habitat analyses (e.g. Titterington et al. 1979), most studies of bird habitats have not taken into account the occurrence of particular tree species. It must be realized, however, that even if tree species were to be included and if correlations were found, there would still be no way of understanding why these tree species are important without detailed observations of what the birds were doing in the different tree species and without measurements of the resources available there.

CONCLUSIONS

A thorough knowledge of the natural history of bird species and their habitat requisites is essential for understanding the relations between birds and their habitats and for determining which habitat components are important to birds. To elucidate further the causal links between birds and their habitats, however, will require comparisons of behavior and habitat responses of the same bird species in habitats that differ in particular ways or that have been manipulated in a manner that will test the importance of some specific habitat parameter(s). Only through such carefully planned observations and experiments can knowledge be obtained that will allow us eventually to predict from habitat data which bird species will occur or not occur in particular habitats. Such a goal is needed if habitats are to be managed scientifically.

LITERATURE CITED

Balda, R.P. 1969. Foliage use by birds of the oak-juniper woodland and ponderosa pine forest in south-eastern Arizona. Condor 71:399-412.

Bhattacharyya, H. 1981. Theory and application of factor analysis and principal components. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Cooley, W.W., and P.R. Lohnes. 1971. Multivariate data analysis. 364 p. John Wiley and Sons, New York, N.Y.

Connor, R.N. 1978. Snag management for cavity nesting birds. p. 120-138. In R.M. DeGraaf, techincal coordinator. Management of southern forests for nongame birds: Proceedings of a workshop [Atlanta, Ga., Jan. 24-26, 1978]. USDA Forest Service General Technical Report SE-14, 176 p. Southeastern Forest Experiment Station, Ashville, N.C.

Franzreb, K.E. 1978. Tree species used by birds in logged and unlogged mixed-coniferous forests. Wilson Bulletin 90:221-238.

Hilden, O. 1965. Habitat selection in birds. Annales Zoologici Fennici 2:53-75.

Holmes, R.T., R.E. Bonney, Jr., and S.W. Pacala. 1979a. Guild structure of the Hubbard Brook bird community: a multivariate approach. Ecology 60:512-520.

Holmes, R.T., J.C. Schultz, and P. Nothnagle. 1979b. Bird predation on forest insects: an exclosure experiment. Science 206:462-463.

Holmes, R.T., and S.K. Robinson. 1981. Tree species preferences of foraging insectivorous birds in a northern hardwoods forest. Oecologia 48:31-35.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

James, F.C., and H.H. Shugart, Jr. 1970. A quantitative method of habitat description. Audubon Field-Notes 24:727-736.

MacArthur, R.H., and J.W. MacArthur. 1961. On bird species diversity. Ecology 42:594-598.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an ozark watershed. Ecology 58:810-819.

Sneath, P.H.A., and R.R. Sokal. 1973. Numerical taxonomy. 573 p. W.H. Freeman, San Francisco, Calif.

Titterington, R.W., H.S. Crawford, and B.N. Burgason. 1979. Songbird responses to commercial clear-cutting in Maine spruce-fir forests. Journal of Wildlife Management 43:602-609.

Whitmore, R.C. 1975. Habitat ordination of passerine birds of the Virgin River Valley, Southwestern Utah. Wilson Bulletin 87:65-74.

Whitmore, R.C. 1977. Habitat partitioning in a community of passerine birds. Wilson Bulletin 89:253-265.

Willson, M.F. 1974. Avian community organization and habitat structure. Ecology 55:1017-1029.

APPLIED ASPECTS OF CHOOSING VARIABLES IN STUDIES OF BIRD HABITATS[1]

Robert C. Whitmore[2]

Abstract.--This paper considers the applied aspects of choosing and using habitat variables; aspects that deal with management decisions rather than evolutionary events that shaped observed community patterns. Vegetation structure is one of the central parameters used in the explanation of observed habitat use patterns of birds. When choosing variables one should consider the range of habitats being studied, the nature of the bird species, the practicality of measurement, the time (cost) needed to measure and the biological importance of each variable. Once the variables are chosen, experimental designs by Martinka (1972) and Whitmore (1981) could be used to gather information needed in making management decisions. The designs basically involve comparing bird territories with adjacent areas that are not occupied (nonterritories).

Key words: Blue grouse; discriminant analysis; grasshopper sparrow; habitat management; variables.

INTRODUCTION

Criteria for determining which habitat variables to measure generally fall into two broad categories, theoretical considerations and applied considerations. The former deal with studies that ask why certain species of birds inhabit a given locale and what events shape this process, while the latter deal primarily with management decisions. The objective of this paper is to discuss applied aspects of the question.

When trying to quantify observed habitat use patterns of birds, variables most often measured reflect vegetation structure. There are two readily apparent reasons for this. First, vegetation structure is a central factor determining the habitat that a bird selects (fig. 1) and, although only a proximate factor (Hilden 1965), does evoke the settling response of a bird arriving in spring. Second, between-site differences in vegetation structure often are obvious and while data sets may be voluminous they are relatively easy and inexpensive to obtain, especially when compared to other parameters such as food availability, microclimate and biological interactions. With increasing use of multivariate statistics in habitat studies, the number and complexity of variables measured have increased (James 1971, Anderson and Shugart 1974, Whitmore 1975, 1977). Moreover, it is likely that the number of these types of studies will increase and, in fact, the call for more has already gone out: "Detailed investigation of the relationships among numerous habitat variables and bird populations...should be encouraged. Such investigations will likely provide widely applicable results in a minimum of time" (Verner 1975).

---

[1]Paper No. 1643 of the West Virginia University Agriculture and Forestry Experiment Station. Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Associate Professor, Division of Forestry, Wildlife Biology Section, West Virginia University, Morgantown, WV 26506.

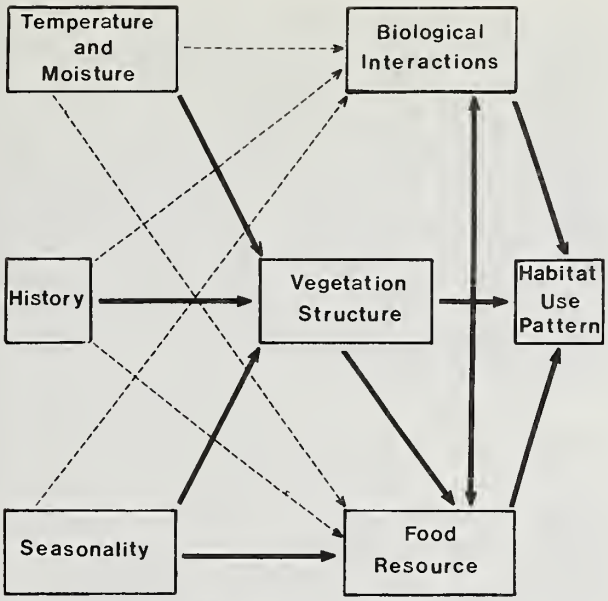


Figure 1. Schematic representation of the relationship between vegetation structure, other variables and observed habitat use patterns in birds. Dashed lines represent minor effects while solid lines represent major effects. Adapted from Karr (1980).

Although the ecologist may not be seeing or measuring what the bird is actually selecting (i.e., "the ecologist may be recognizing distinct habitats or positions along environmental gradients, but the bird species present may not be capable of the same distinctions or their distinctions may not be equivalent to those of the observer", Whitmore 1977), detailed multivariate analysis of species distributions along complex habitat gradients may have predictive value in determining which species will occupy a given site, and may be useful in making management decisions (James 1971, Martinka 1972, Whitmore 1981). However, as pointed out in numerous studies, the choices and number of variables (James 1971) as well as time of year in which they were measured (Whitmore 1979) affects results. Therefore, any multivariate study is only as good as the input variables. The following sections of this paper will first examine criteria for selecting variables and then give an example as to how they might be used in an applied situation.

CRITERIA FOR SELECTING VARIABLES

The plethora of avian habitat studies can be placed along a continuum of "quantitativeness" ranging from subjective, visual analysis (extensive), such as the Missouri Plan for habitat evaluation, to microhabitat analysis (intensive) which may be as detailed as counting the number of grass stems in a 5 ha field. As an example of an extremely extensive methodology I am reminded of the waterfowl manager who developed a three category system for analyzing marsh vegetation:

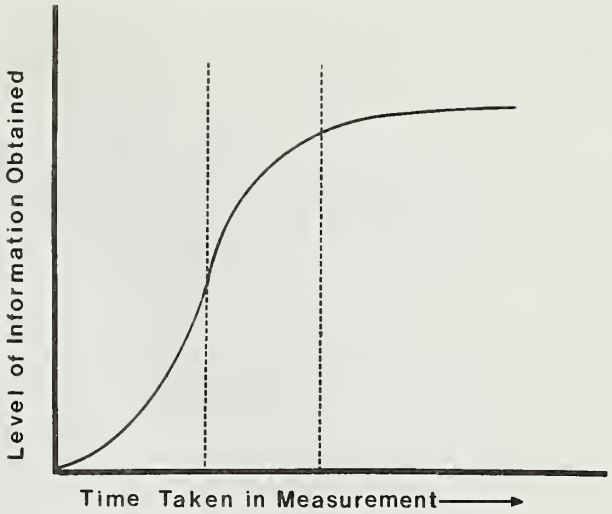


Figure 2. A curve representing the relationship between the level of information obtained and the time taken in measurement. Dashed zone indicates an optimal area where the ratio produces the most benefit per unit expended.

If a duck can swim through it in a straight line--open; weaving line--moderate; can't make it through--dense.[3] In the middle of the continuum of quantitativeness lies what I term "quick and dirty" methods of habitat analysis. One of the most widely used of such techniques was proposed by James and Shugart (1970) and has been used in a variety of studies (see Whitmore 1975, 1977). More detailed analysis, the intensive methods, can be seen in works of Cody (1968) and Wiens (1969) for their analyses of grassland habitats.

A fundamental criterion for picking a sampling technique is the time needed and relative cost of each method. When plotting the time taken in measurements versus the level of information obtained a Gaussian curve emerges (fig. 2). Somewhere in the middle of the curve is a "zone of optimality" where a maximum return in information for amount of time (money) expended is obtained. An effective sampling scheme should fall within this zone. For the scope of many studies the James and Shugart (1970) technique fits the bill. It remains for others to develop techniques that will shift the curve to the left on the abscissa thereby increasing the ratio of information/time spent.

The nature of variables measured depend on the range of habitats being considered. Quantifying habitat use of birds in habitats ranging from grassland to forest may require less intensive variables than are necessary to separate four Panamanian forest plots. Another consideration is what kind of species are you

[3] Personal communication with E.D. Michael, Professor, West Virginia University.

dealing with? Would variables used to quantify habitats of an acorn woodpecker (Melanerpes formicivorous) also be appropriate for a grasshopper sparrow (Ammodramus savannarum)?

Some practical aspects when deciding which variables to pick might fall into three general categories. First, variables should be measurable to the desired level of precision. Second, variables should be biologically meaningful. For example, it may be possible to measure tree roots to 60 cm away from the trunk below the surface of the ground, but what possible meaning could this have to a canopy foraging bird such as a cerulean warbler (Dendroica cerulea)? Third, variables should be relevant to the species in question. Would percent grass cover or the ratio of grasses to forbs be important to a bark forager such as a brown creeper (Certhia familiaris)?

As a preface to the next section I would like to make two comments that seem to have a bearing on many of the papers in these proceedings. First, we scientists are often guilty of measuring everything there is to measure, trusting stepwise discriminant analysis or some other technique to sort things out, simply because we don't know what is important. This may be viewed simply as a substitutive for thinking. And second, if you miss the key parameters you can go out and measure everything else, use the most sophisticated multivariate techniques and still have nothing of biological importance. However, if one is careful in picking variables and follows, at least generally, the concepts listed above, then multivariate statistics may be useful in sorting out habitats and making management decisions.

USING MULTIVARIATE STATISTICS IN WILDLIFE MANAGEMENT

Assuming correct variables are measured, this section deals with application of multi-variable data sets to wildlife management. Standard management questions might include:

1. What effect will habitat alteration have on the bird species in question (BSIQ)?
2. What structural characteristics are found in habitat type A but not in type B that allow the BSIQ to live in the former but not the latter?
3. How can I make habitat type B suitable for the BSIQ?
4. How can I get more of the BSIQ into type A?
5. Can I introduce the BSIQ into habitat type C?

From reading papers in the Journal of Wildlife Management it is apparent that multivariate statistics as a tool in making management decisions have not had widespread use. The only multivariate paper of interest that I could find in recent volumes of the Journal of Wildlife Management presented a technique for habitat analysis that I feel has broad application in wildlife management. However, it has apparently been overlooked. Martinka (1972) published a study comparing habitats of blue grouse (Dendragapus obscurus) with surrounding habitat that appeared similar but did not have blue grouse using it. He termed the latter sites "nonterritories." This type of analysis, territory vs. nonterritory, seems appropriate in addressing several questions listed above. Basically, Martinka found that by using discriminant analysis he could correctly classify 96% of his plots as either territories or nonterritories and by so doing develop a model for management recommendations.

This same line of reasoning may be of fundamental use in documenting avian habitat use in general. Most studies are designed to compare different bird species or different communities. Perhaps it would be more fruitful to look at a single species, comparing sites in which it is found with similar habitat where it is not found. As an example of this method I summarize a paper on grasshopper sparrows (Whitmore 1981). This species is a common breeder on reclaimed surface mine grasslands in northern West Virginia, yet not all sites have grasshopper sparrows and on those that do it is rare for the entire area to be used. Questions asked were 1) are there vegetational structure characteristics that are identifiably different between used and unused areas, and 2) if there are, can the unused areas be managed to make them usable? On one reclaimed site in Preston County, West Virginia all grasshopper sparrow territories were located and mapped using the territory flush technique (Wiens 1969). Once mapped, vegetation structure variables were measured in each territory. All areas on the same mine that were not used by grasshopper sparrows (nonterritories) were located and sampled. By using stepwise discriminant analysis it was found that territories could be separated from nonterritories with 100% accuracy on a vegetation density gradient. Nonterritories had significantly greater ($P < 0.01$) values for most grass, shrub and litter cover variables and significantly lower ($P < 0.01$) percent bare ground. A change in the habitat from unusable to usuable could be manifested by any of a variety of range management techniques that would reduce vegetation density and litter build up; the easiest being fire. Manipulation experiments are currently being designed in an attempt to shift unused habitat to usable.

Further extension of the above two examples may be necessary to fit different species or groups of species. However, the basic concept could be used in a variety of wildlife studies ranging over many habitat types.

SUMMARY

Many studies presented in the current body of literature are based on a set of variables that seem to be picked with little regard for the habitats or species in question and more often

than not are based on either convenience or "what someone else has done." This paper presents a set of ideas that should be considered before initiating a bird habitat study. Two examples of an experimental design based on vegetation analysis of territories vs. nonterritories, one with a game bird the other with a songbird, show great potential for future use in making management decisions. Scientists should remember that the choice of vegetation variables, the number of variables, and the time required for measurement can greatly affect results in a multivariate study. Perhaps the greatest amount of time should be spent in the design and variable selection stages of the experiment rather than in the collection of data.

LITERATURE CITED

Anderson, S.H., and H.H. Shugart. 1974. Habitat selection of breeding birds in an east Tennessee deciduous forest. Ecology 55:828-837.

Cody, M.L. 1968. On the methods of resource division in grassland bird communities. American Naturalist 102:107-147.

Hildén, O. 1965. Habitat selection in birds. Annales Zoological Fennici 2:53-75.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

James, F.C., and H.H. Shugart. 1970. A quantitative method of habitat description. Audubon Field-Notes 24:727-736.

Karr, J.R. 1980. Geographical variation in the avifaunas of tropical forest undergrowth. Auk 97:283-298.

Martinka, R.P. 1972. Structural characteristics of blue grouse territories in southwestern Montana. Journal of Wildlife Management 36:498-510.

Verner, J. 1975. Avian behavior and habitat management. p. 39-58. In Smith, D.R., technical coordinator. Management of forest and range habitats for nongame birds: Proceedings of a symposium [Tucson, Ariz., May 6-9, 1975]. USDA Forest Service General Technical Report WO-1. 343 p. Washington, D.C.

Whitmore, R.C. 1975. Habitat ordination of the passerine birds of the Virgin River Valley, southwestern Utah. Wilson Bulletin 87:65-74.

Whitmore, R.C. 1977. Habitat partitioning in a community of passerine birds. Wilson Bulletin 89:253-265.

Whitmore, R.C. 1979. Temporal variation in the selected habitats of a guild of grassland sparrows. Wilson Bulletin 91:592-598.

Whitmore, R.C. 1981. Structural characteristics of grasshopper sparrow habitat. Journal of Wildlife Management 45 (In press).

Wiens, J.A. 1969. An approach to the study of ecological relationships among grassland birds. 93p. Ornithological Monographs No. 8, American Ornithologists Union.

DISCUSSION

JAMES DUNN: How did you decide on the locations of the non-territories?

BOB WHITMORE: We had the entire surface mine mapped as to territory location. All areas that did not have territories were sampled, subject to the constraint that an entire 50 m radius circle would fit within it.

JIM WOEHR: Do you agree that vegetation patchiness is an important habitat variable, and if so, how should we measure it and how can we statistically test for differences in patchiness?

BOB WHITMORE: I agree it is important, but I do not have a real handle on how to measure it.

JIM WOEHR: In comparing territories with non-territories, do you think it is critical to study a habitat over several years to look at fluctuations in bird density in different habitats? It seems to me that the "best" habitat will always have more birds, whereas marginal habitats will have fewer birds, or none at all, in years of low populations.

BOB WHITMORE: I totally agree that in most natural systems yearly changes in environmental factors such as climate will affect avian densities and habitat use patterns, often forcing birds into "suboptimal" habitats. Although it is difficult to determine quantitatively, I feel that the strip-mine birds, as yet, are not habitat limited and, therefore, are taking the best available. Newly reclaimed surface mine habitat is being created faster than the birds are capable of using it. We definitely need some bad years to study.

TECHNIQUES FOR SAMPLING AVIAN HABITATS[1]

Barry R. Noon[2]

Abstract.--Standardized methodologies for the sampling of bird-related vegetation structure in forest and non-forest habitats are proposed. For forest habitats, the methodology is based largely on techniques proposed by James and Shugart (1970). For non-forest habitats, the methodology is a synthesis of many previously published techniques, particularly those of Wiens (1969). For each habitat type a detailed sampling protocol and sample field data sheet are provided. In addition, statistical and biological considerations for the location of sampling points are discussed. The paper ends with an argument in favor of standardized methods of sampling avian habitats.

Key words: Avian; habitat structure; sampling techniques; standardized procedures.

INTRODUCTION

There are two common goals when sampling avian habitat structure. The first is to measure features of the habitat that will allow an accurate determination of the species' habitat requirements. Habitat parameters believed to be at least proximally related to a species' survivorship and reproductive success in that habitat are selected for measurement. The second is the ability to make accurate predictions of a species' response to habitat change and to anticipate possible detrimental effects to a species population from various land-use practices. This second goal is contingent upon having achieved the first.

To date, there have been a variety of approaches to describing the habitat associations of bird species (Niemi and Pfannmuler 1979). These include the non-quantitative successional stage (Kendeigh 1945, Johnson and Odum 1956, Martin 1960, Haapanen 1965, Holt 1974) and life form (Thomas et al. 1975, 1976) approaches, as well as quantitative approaches based on statistically defined species-habitat associations (e.g., Rotenberry and Wiens 1978, Pfannmuller 1979) or the multidimensional habitat-niche approach adopted from Hutchinson (1957). The statistically based, quantitative approaches are currently favored because they are generally less subjective and yield more information on specific niche requirements than do non-quantitative approaches.

A holistic, community based approach to species-habitat associations has been outlined by Niemi and Pfannmuller (1979). The procedure is based on a cluster algorithm which groups avian communities, representing various points along a habitat gradient, into hierarchial clusters on the basis of their similarity in species composition. The researcher investigates whether the groups formed by the clustering algorithm(s) suggest any structural features of the habitat gradient recognized as important by the birds. The same community census data can be subjected to an inverse clustering algorithm which groups together bird species that show similar distributions across the habitat gradient (Pfannmuller 1979).

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Research Biologist, Migratory Bird and Habitat Research Laboratory, U.S. Fish and Wildlife Service, Laurel MD 20811.

The distribution of species associations and the community clusters may yield insights into the habitat requirements of individual species and also indicate which habitats are recognized as distinct types.

The avian community approach does not yield information on specific habitat niche requirements of individual species. However, it is a valuable preliminary to more detailed niche analysis because it accurately delimits the range of habitats to be sampled for a particular species or species association.

To clarify the relationship between birds and habitat structure, we must develop techniques of habitat measurement which fulfill the following criteria:[3]

1. Yield efficient, precise, and accurate estimates of the habitat parameters of interest.
2. Yield data amenable to various statistical tests, especially in multivariate analyses on which predictive models would subsequently be based.
3. Quantify attributes of the habitat that are biologically relevant to the species (and biologically interpretable) and that can be unambiguously communicated to other researchers.
4. Yield data whose predictive capabilities can be tested by simulated tests.
5. Are applicable in different structural habitat types.

The sampling protocol that follows is based on a multivariate approach to habitat selection, specifically on the niche concept as formalized by Hutchinson (1957). The rationale for employing these techniques is the belief that a species' response to habitat structure is not univariate. That is, the suitability of a habitat patch to an individual bird is a function of several interrelated habitat parameters whose combined effect (in a mathematical sense) determines the habitat's suitability.

When initiating a habitat-based niche analysis of one or more species, three major aspects of the experimental design must be resolved before sampling begins. I will address each of these below.

How Finely Should Habitat Be Measured?

Obviously, the habitat structure must be measured in enough detail so that factors believed important to the species being studied are accurately estimated. In addition, habitat structure should be measured so as to bring into focus differences that may allow two similar bird species to coexist. These data will also give valuable insights into community organization.

As a general rule, the more apparently homogeneous the habitat, the finer it will need to be sampled in order to detect its inherent heterogeneity. Thus, grassland habitats require sampling vegetation structure at a much finer level than would forest habitats (cf. Wiens 1969 with Whitmore 1975).

Often the researcher is uncertain how finely species are discriminating habitat and, as a consequence, feels the need to sample both micro- and macrohabitat gradients. Macrohabitat descriptions require sampling a relatively large area, while sampling microhabitat gradients over the same area would result in prohibitive time constraints. A workable solution is to use nested samples with microhabitat variables estimated from sampling areas contained within the larger sampling units used to estimate macrohabitat variables. Titterington et al. (1979) provides a good example of the use of nested vegetation sampling plots in the study of avian populations.

How Many Samples Should Be Taken?

This topic is covered in more detail by Johnson (1981); however, a brief point follows. A sampling criterion such as being within 10 percent of the population mean 90 percent of the time should be established for each habitat parameter estimated. Sample size formulas to meet these and other criteria are given in Cochran (1963) and Steel and Torrie (1960). Unfortunately, for some parameters of interest the appropriate sample size to meet these criteria is impossible to attain (Noon, unpublished data). This is particularly true for habitat variables or gradients that are very patchily distributed in the study area (e.g., the percent of coniferous vegetation in different vertical strata in a primarily deciduous woods). For example, using the above criteria, a mature deciduous forest plot in New York State (Noon 1979) required over 100, 0.04 ha circular plots to estimate most structural habitat parameters. Changing the criteria so as to be within 10 percent of the mean 80 percent of the time lowered the requisite sample size for the same variables to approximately 60 circular plots.

Where Should Samples Be Taken?

A major assumption of most statistical models is that the statistics themselves have been estimated from a random sample of the population under study. When estimating structural parameters of the vegetation relative to the bird community, stratified random sampling may be more appropriate than pure random sampling. Stratification will result in a more uniform sampling of the study area. I suggest that the stratification be based on the following criteria:

1. Along obvious lines of habitat

[3]Personal communication with Frances James, Associate Professor, Department of Biological Science, Florida State University.

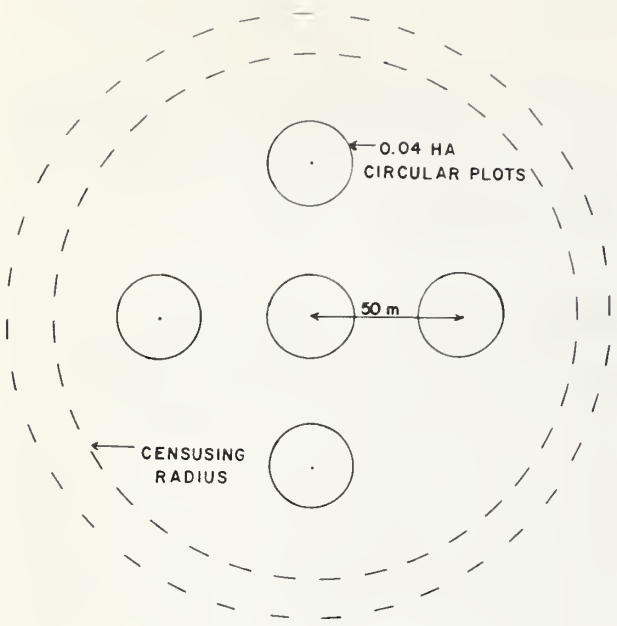


Figure 1. Point-count census technique with variable censusing radius. Five, 0.04-ha circular plots for sampling vegetation structure are clustered about the censusing point.

heterogeneity within the study area. It is very important that patches of structurally distinct habitat be represented in data from which estimates of available habitat are made.

2. By the location of individual birds in the habitat; that is, let the bird define the sampling location.

The importance of bird-defined sampling locations is clarified by examining how easily erroneous inferences about a species' habitat requirements arise from improper sampling. I will illustrate the problem with bird census data collected with the point count technique (Ferry and Frochot 1970) and habitat data subsequently gathered from 0.04-ha (0.1 acre) circular plots (James and Shugart 1970) centered around the censusing point.

Figure 1 illustrates a sampling design with five, 0.04-ha circles clustered about the censusing point. A hypothetical study using this experimental design would conduct many point-counts along a habitat gradient and associate all the species recorded at a particular censusing location with the mean habitat vector (averaged across the five circles) at that location. A two-group discriminant function analysis may then be conducted, for each species, to distinguish points where the species was found from points where it was not detected. Discrimination of these "present" and "absent" groups would be interpreted in terms of differences in habitat structure.

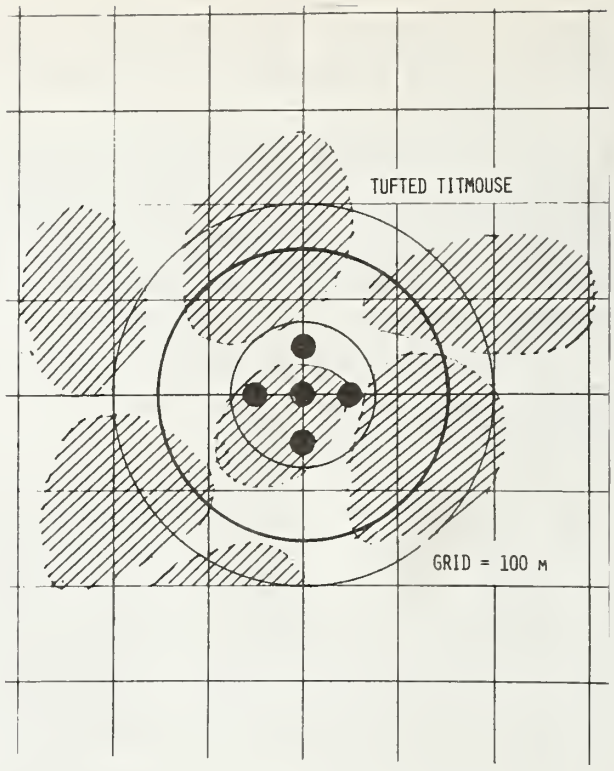


Figure 2. As in fig. 1 but with actual tufted titmouse territories (determined by territory mapping) detectable from the point-center positioned relative to the vegetation samples.

Figure 2 illustrates the actual location of tufted titmouse territories around one of these census points. Note that habitat data from non-utilized areas are associated with the species. If the habitat sampled from non-utilized areas is substantially different from the utilized areas, then it will be quite difficult to discern the true habitat preferences of a species with this experimental design. The problem is not unique to this sampling situation but may arise anytime habitat data associated with a particular individual represent areas not actually utilized by that individual.

METHODS

I have outlined a sampling methodology that combines random, or statified random, sampling with bird defined sampling locations as follows (see figs. 3 and 4): 1) Samples of vegetation structure are taken at randomly selected locations within the study plot boundaries. Vegetational characteristics of individual territories are determined by superimposing a map of the territory boundaries over a map of the numbered sampling sites and summing the data from all sample sites included within the boundaries (Wiens 1969). Only samples falling totally within territory boundaries are assigned to that species. 2) Territories that by chance were not sampled

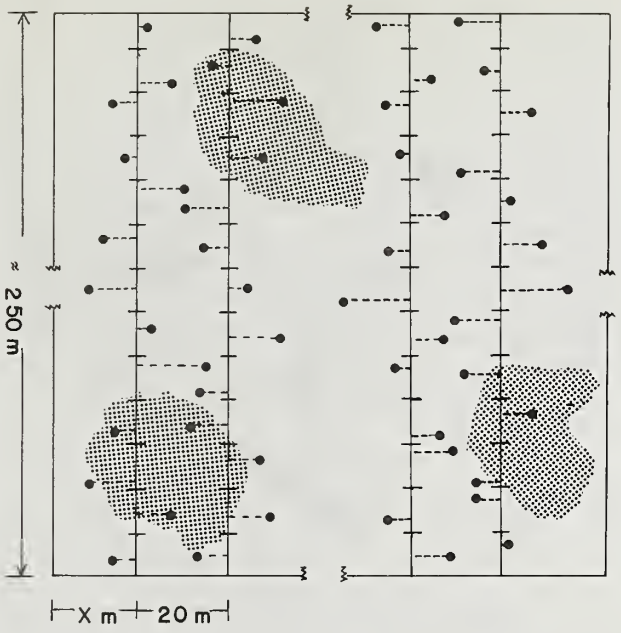


Figure 3. Line transect system with randomly located sampling points indicated. Territories are superimposed over sample locations.

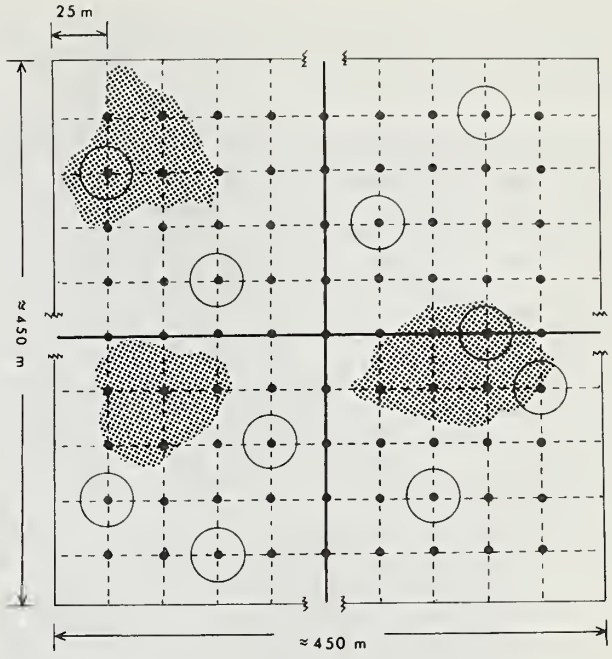


Figure 4. Areal plot system with randomly located 0.04-ha circular plots indicated. Territories are superimposed over sample locations. The 25-m grid system illustrated here would only be necessary in very closed habitats.

are sampled in some random fashion.

Given sufficient data, this sampling scheme allows direct comparison between species' structural habitats as well as a comparison of each species against a random sample of the available habitat. The sampling protocol is amenable to studies in both non-forest (≤ 25 percent cover by trees) and forest habitats. I suggest that non-forest habitats be sampled by line transect methods and forest habitats by a modification of the James-Shugart (1970) 0.04-ha circular plot technique. Transects should be placed across contour lines (Oosting 1956), and evenly spaced across the study plot with the location of the initial transect line determined by some random process (fig. 3). For forest habitats, sampling points can be determined by a random (or stratified random) selection of grid points used in the territory mapping procedure (fig. 4).

Sampling Protocol

Non-forest Habitats

Avian habitat studies in "non-forest" environments (≤ 25 percent cover by trees) have used primarily transect sampling procedures to establish sampling locations. Along the transect, or at randomly selected points, vegetation structure has been estimated by fixed quadrat, line intercept, or point-centered quarter techniques (refer to Smith 1974 for a general description of these procedures). I propose a sampling protocol that combines line-intercept and point-centered quarter techniques. In developing this protocol I have borrowed extensively from the work of others, particularly Wiens (1969).

Line-intercept Variables. The line is considered to be a belt 1 cm wide running along one side of the transect. The transect is divided (stratified) into 10-m intervals. Within each interval the distance on the transect line interrupted by specific life forms or habitat features is recorded (appendix IA).

For shrub and tree life forms, coverage is estimated by the downward vertical projection of their foliage lying above the line. Within each 10-m segment the minimum total amount of coverage summed over all habitat features is equal to the length of that segment; however, total coverage is usually higher. Data collected in this way allow calculation of the frequency, density, and dominance of each habitat feature (Smith 1974).

If the ground-level vegetation is not arrayed as discrete patches then line-intercept techniques are very difficult to use. When the ground cover is an intricate mosaic, line-intercept techniques should be replaced by point-intercept methods. The procedure is to uniformly select numerous points within each 10-m interval and record the presence and absence of specific ground cover life forms intercepted at these points.

Statistical treatment of the results and cautionary notes on methods of point sampling are found in Goodall (1952).

Point-quarter Variables. Each 10-m interval along the transect will correspond to a sampling unit. Within each interval the sampling point is determined by selecting three random digits to indicate 1) the linear distance in meters along the transect interval given by the first digit, 2) the side of the line to be sampled (odd digit = left; even digit = right) given by the second digit, and 3) the number of 0.5-m intervals to be marked off perpendicular to the transect given by the third digit. At the sampling point, quarters are established by placing two 1-m sticks on top of each other oriented in the cardinal directions to form a '+'. Within each quarter the following variables are estimated:

A. Species, distance to, and height of the nearest shrub (woody vegetation > 1 m tall and < 3 cm dbh).
B. Species, distance to, dbh, and height of the nearest sapling (3 cm ≤ dbh < 8 cm) and tree (≥ 8 cm dbh).
C. Vertical vegetation density. At each of the four ends of the meter sticks, vertically lower into the vegetation a 1-cm diameter rod graduated into the following intervals: 0-0.3 m, 0.3-1 m, and 1-2 m. Vertical vegetation density is estimated by recording the number of contacts of vegetation falling within each interval as well as visually estimating the number of contacts from 2-9 m and > 9 m.
D. "Effective vegetation height". At the approximate intersection point of the two meter sticks record "effective vegetation height" as detailed by Wiens (1969).

Methods to calculate density, dominance, and frequency estimates from point-quarter data are given in Smith (1974). A sample field data sheet is given in appendix II.

Forest Habitats

The sampling protocol for forest habitats (> 25 percent cover by trees) is based on the James and Shugart (1970) 0.04-ha (0.1 acre) circular plot techniques. I have modified the techniques to include additional data and to clarify existing ambiguities (cf. James 1978); however, the core of the technique remains unaltered. A sample field data sheet for recording the estimates outlined below is given in appendix III.

The following habitat features are measured within the 0.04-ha (radius = 11.3 m) circle:

1. The diameter at breast height (dbh), 1.3 m above the ground, of all saplings and all standing trees. The dbh values will be recorded by tree species within nine size classes (appendix IB). Standing, dead trees are recorded separately by size class. The size classes are almost identical to those proposed by James and Shugart (rounded to nearest cm) but with the addition of a smaller dbh size class, S (sapling: 3-8 cm dbh).
2. Shrub density at breast height is estimated along two transects running in the cardinal directions and centered within the 0.04-ha circle (fig. 5). The observer proceeds along the transect lines counting the number of woody stems < 3 cm dbh intersected with his body and outstretched arms at breast height. Counted stems only include the main stem and those stems branching from the main stem below breast height. The total number of contacts made in two transects (each 22.6 m long) times 125 is used to give an estimate of the number of shrub stems per ha. The contribution of deciduous and coniferous shrubs is recorded separately.
3. Canopy cover and ground cover are estimated by sighting through an ocular tube, made from a cardboard cylinder with cross hairs at one end. The observer walks along the transect lines used to estimate shrub density sighting up to the canopy and recording a total of 20 (10 each transect) plus or minus readings indicating the presence or absence,

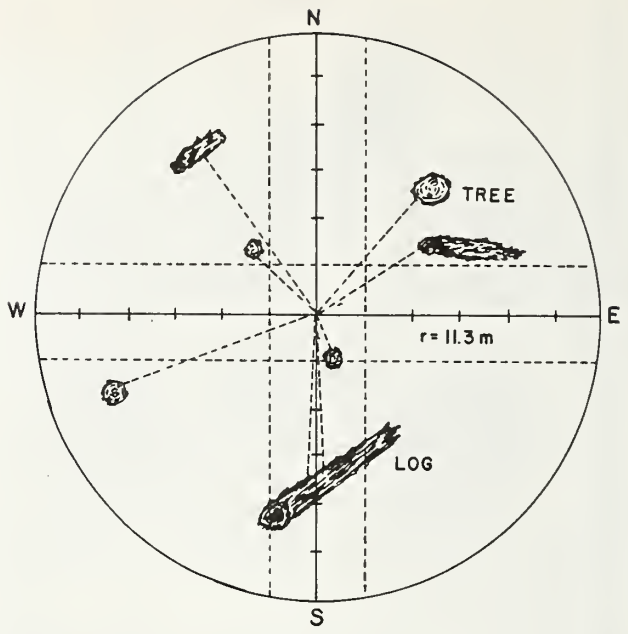


Figure 5. Circular plot (r = 11.3 m) used to estimate vegetation structure in forest habitats. Rectangular plots used to estimate shrub density, transects establishing quarters, point-quarter tree and log variables, and sample points for estimating canopy and ground cover are all indicated.

respectively, of green vegetation at the intersection point of the cross hairs. Percent of the canopy cover contributed by coniferous foliage is recorded in addition to total canopy cover. Green vegetation within a meter of the ground is recorded in an identical manner except that the observer sights downward through the tube held at waist height (approximately 1 m above the ground). Canopy and ground cover are recorded as percents (i.e., number of hits/20 x 100).

4. A qualitative plant dispersion index is recorded for ground (0-1 m tall) and shrub (> 1 m tall and < 3 cm dbh) strata plants. The index is identical to that proposed by Emlen (1956). The categories are
   E - Even matrix (more or less randomly dispersed)
   I - Irregular or uneven (indistinct clumps)
   SC - Small clumps
   LC - Large clumps
   SR - Small distinct rows or hedges
   LR - Large distinct rows or strips.
5. Canopy height should express the average height (m) of the canopy within the 0.04-ha circle. The observer should make several measurements (with a clinometer, range finder, Abney level, or similar instrument), average these measurements, and record this average. Also, the maximum and minimum estimates of canopy height are recorded.
6. Slope is estimated with the aid of a clinometer. This estimate is the maximum slope within the circular plot.
7. Indices of tree and log dispersion are gathered from point-quarter techniques. The point is centered in the circle and the quarters are established by the transects used to estimate shrub density (fig. 5). Within each quarter the distance to, and dbh size-class of the nearest tree are recorded. In addition, the distance to, total length of, and dbh size-class of the largest (by diameter) fallen log ($\geq$ 1.5 m in length and $\geq$ 8 cm dbh) are recorded (fig. 5). The dbh size-class of the log is determined by the maximum dbh attained throughout its length whether lying totally within the plot or not.
8. Understory foliage volume is estimated with a density board (Wight 1938, DeVos and Mosby 1969). The density board, or drop cloth, (fig. 6) is divided into four height intervals, 0-0.3 m, 0.3-1 m, 1-2 m, and 2-3 m, corresponding to low ground, high ground, and low and high shrub levels, respectively. Foliage volume contributed by the sapling level is indirectly assessed by the number of trees falling in dbh size-class S. The drop cloth is placed at each of the four points where the transect lines intersect the edge of the circle. Four readings

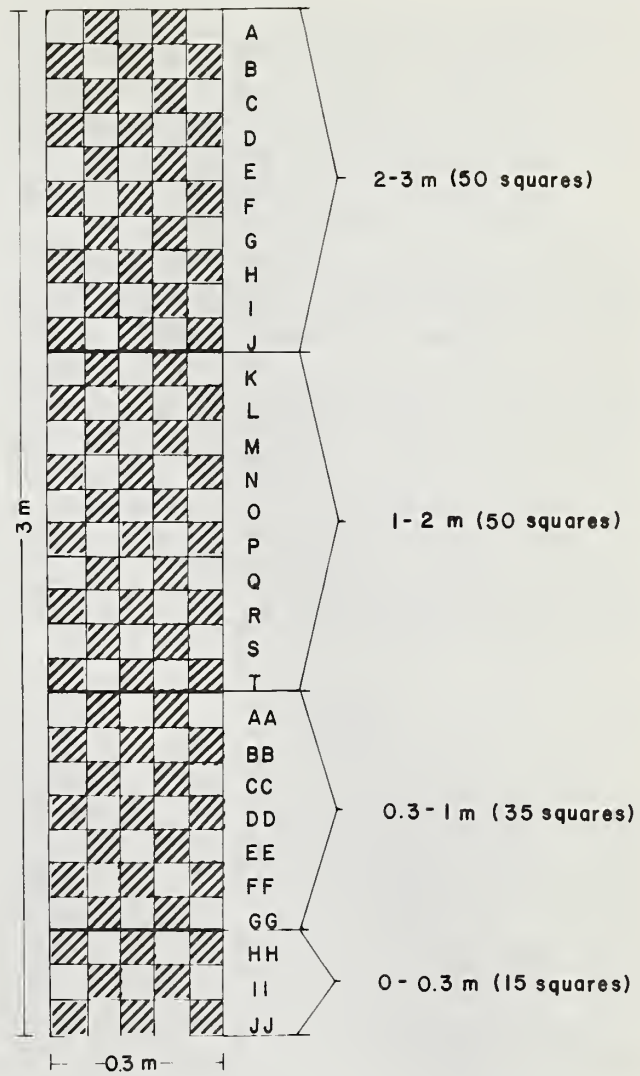


Figure 6. Density board ("drop cloth") used to estimate foliage volume from 0-3 m.

are made from the center of the circle (i.e., 11.3 m distance) sighting along each of the transects (i.e., N,S,E,W). The observer counts the number of squares within each height interval at least 50 percent obscured by foliage and records this number. To minimize parallax problems, foliage volume in the first two height intervals is estimated from a crouching position, and from a standing position for the upper two intervals. Problems may arise if dense, low vegetation lies within the immediate vicinity of the center of the circle ($\leq$ 1.5 m). In this situation the observer should move to the side the minimum distance necessary to give an unobstructed view within the first 1.5 m.

9. Dominant shrub species and ground cover (ground to 1 m tall) life forms are recorded in rank order with the most common species, or life form, listed first. The ranking is estimated only within the 0.04-ha circle. A list of ground cover life forms to discriminate is given in appendix IC.

Equipment needed for estimating density, basal area, and frequency of trees, canopy height, shrub density, and percent ground and canopy cover is given in James and Shugart (1970). Sapling trees are measured with a forester's diameter tape. Density boards (drop cloths) can be made from a variety of materials but those made with either oil cloth or vinyl prove to be both resilient and portable. The gradations are scored on the material with an indelible marking pen. The drop cloth is extended to its full height with the aid of a "telescoping" aluminum pole such as is used by painters and window washers. When not in use, the drop cloth may be rolled around the wooden dowel used to attach the top of the cloth to the aluminum pole.

DISCUSSION

Many of the papers presented in these proceedings have used a habitat sampling protocol similar to that outlined here and illustrate a variety of approaches to data analysis and interpretation. Additional sources of references are Anderson and Shugart 1974; Bertin 1977; Bourgeois 1977; Cody 1968, 1978; Cody and Walter 1976; Conner and Adkisson 1976; James 1971; Karr 1968, 1971, 1976; Karr and Roth 1971; Noon 1981; Noon and Able 1978; Rabe 1977; Rice 1978; Roth 1976; Smith 1977; Sturman 1968; Titterington et al. 1979; Whitmore 1975, 1977, 1979; and Wiens 1968, 1973, 1974.

Ecologically meaningful information may be derived from understory foliage volume and tree size-class data. Vertical foliage volume is directly estimated up to 3 m by the drop cloth. Further, indirect estimates of foliage volume may be extracted from dbh size-class data for specific species or types of trees (e.g., hardwoods and conifers; Harris et al. 1973, Weinstein[4], Smith[5]). This results from the predictable relationship between tree height and dbh for the lower size classes (S-C) of most species of trees (fig. 7; e.g., Curtis 1967; Schreuder and Hafley 1977). Coupling these two sources of information by tree species or tree type allows estimates of foliage volume by vertical strata.

As an example, consider the data collected by

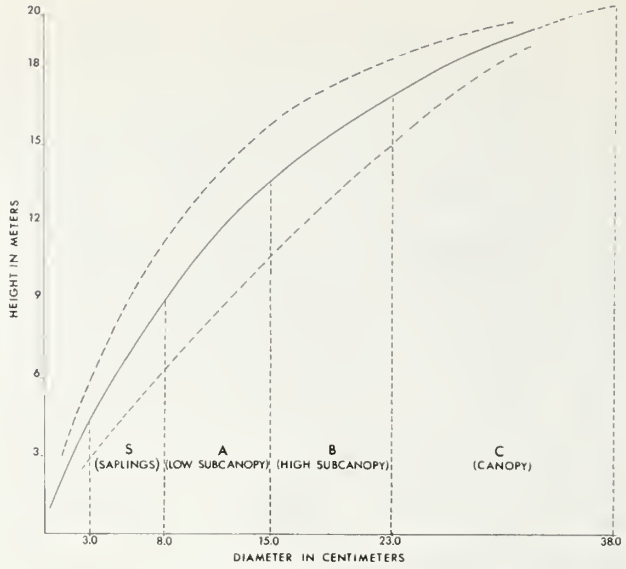


Figure 7. Regression of tree height on diameter at breast height (modified from Schreuder and Hafley 1977).

the modified James-Shugart techniques outlined here. Estimates of foliage volume from the drop cloth are stratified as 0-.3 m (low ground), 0.3-1 m (high ground), 1-2 m (low shrub), 2-3 m (high shrub), and for the tree strata (fig. 7) as 4-9 m (saplings), 9-14 m (low subcanopy), 14-17 m (high subcanopy), and > 17 m (canopy). Estimates of foliage volume at different vertical strata allow calculation of total foliage volume as well as both vertical and horizontal foliage distribution and heterogeneity.

Several horizontal habitat heterogeneity indices have been proposed (Wiens 1974, Roth 1976, Anderson et al. 1979). However, the technique used by Anderson et al., based on the variance in foliage volume both within and across vertical strata, is most amenable to data of the type considered here. Insights into the relationship between total foliage volume and bird species abundance, and between vertical and horizontal foliage heterogeneity and bird species diversity may arise when detailed vegetation data are combined with information on the spatial distribution of birds in the habitat.

CONCLUDING REMARKS

The standardized procedures outlined above are not meant to be a constraint on additional or new ways of looking at avian habitat structure. However, I believe that several points can be made in favor of some degree of standardization in methodology. First, standardized methods will clarify communication among researchers and avoid ambiguities in the interpretation of avian-habitat interrelations. Second, standardized data will allow researchers to address questions requiring comparable data sets. For example, two

[4]Manuscript in preparation, D.A. Wienstein, Environmental Science Division, Oak Ridge National Laboratory.

[5]Personal communication with Thomas R. Smith, Environmental Science Division, Oak Ridge National Laboratory.

researchers who had studied the habitat relations of a species in different parts of its range could collaborate to examine the extent of geographical variation in habitat use by that species. The Breeding Bird Census data, with associated James-Shugart structural vegetation data, have already permitted geographical comparisons (Robbins 1978, Noon et al. 1980, James and Wamer, ms) because of standardization in sampling methodology. Finally, once researchers and land managers have reached a common interpretation of the habitat parameters, research findings can be more directly incorporated into land management practices for avian species.

ACKNOWLEDGMENTS

Stanley Anderson, Deanna Dawson, Douglas Inkley, Frances James, and Chandler Robbins have contributed extensively to the ideas presented here. In addition many biologists at the Migratory Bird and Habitat Research Laboratory have field tested these techniques and offered valuable criticism. However, I did not always heed the advice of my colleagues and much of the methodology and emphasis on "important" variables reflects my personal biases.

LITERATURE CITED

Anderson, B.W., R.D. Ohmart, and J. Disano. 1979. Revegetating the riparian floodplain for wildlife. p.318-331. In Johnson, R.R., and J.R. McCormack, technical coordinators. Strategies for protection and management of floodplain wetlands and other riparian ecosystems: Proceedings of a symposium [Callaway Gardens, Ga., December 11-13, 1978]. USDA Forest Service General Technical Report WO-12, 410 p. Washington, D.C.

Anderson, S.H., and H.H. Shugart. 1974. Habitat selection of breeding birds in an east Tennessee deciduous forest. Ecology 55:828-837.

Bertin R.I. 1977. Breeding habitats of the wood thrush and veery. Condor 79:303-311.

Bourgeois, A. 1977. Quantitative analysis of American woodcock nest and brood habitat. Proceedings Woodcock Symposium 6:109-118.

Cochran, W.G. 1963. Sampling techniques. 413 p. John Wiley and Sons, New York, N.Y.

Cody, M.L. 1968. On methods of resource division in grassland bird communities. American Naturalist 102:107-147.

Cody, M.L. 1978. Habitat selection and interspecific territoriality among the sylviid warblers of England and Sweden. Ecological Monographs 48:351-386.

Cody, M.L., and H. Walter. 1976. Habitat selection and interspecific interactions among Mediterranean sylviid warblers. Oikos 27:210-238.

Conner, R.N., and C.S. Adkisson. 1976. Discriminant function analysis: a possible aid in determining the impact of forest management on woodpecker nesting habitat. Forest Science 22:122-127.

Curtis, R.O. 1967. Height-diameter and height-diameter-age equations for second-growth Douglas Fir. Forest Science 13:365-375.

DeVos, A., and H.S. Mosby. 1969. Habitat analysis and evaluation. p. 135-172. In Giles, R.H., Jr., editor. Wildlife management techniques. The Wildlife Society, Washington, D.C.

Emlen, J.T. 1956. A method for describing and comparing avian habitats. Ibis 98:565-576.

Ferry, C., and B. Frochot. 1970. L' avifaune indificatrice d'une foret de chenes pedoncules en bourgogne etude de deux successions ecologique. La Terre et la Vie 24:153-250.

Goodall, D.W. 1952. Some considerations in the use of point quadrats for the analysis of vegetation. Australian Journal of Scientific Research, Series B 5:1-41.

Haapanen, A. 1965. Bird fauna of the Finnish forests in relation to forest succession. I. Annales Zoologici Fennici 2:153-196.

Harris, W.F., R.A. Goldstein, and P. Sollins. 1973. Net above ground production and estimates of standing biomass on Walker Branch Watershed. p. 41-64. In Young, H.E., editor. IUFRO Biomass Studies. University of Maine Press, Orono, Me.

Holt, J. 1974. Bird populations in the hemlock sere on the Highlands plateau, North Carolina, 1946 to 1972. Wilson Bulletin 86:397-406.

Hutchinson, G.E. 1957. Concluding remarks. Cold Spring Harbor Symposium on Quantitative Biology 22:415-427.

James, F.C. 1971. Ordination of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

James, F.C. 1978. On understanding quantitative surveys of vegetation. American Birds 32:18-21.

James, F.C., and H.H. Shugart. 1970. A quantitative method of habitat description. Audubon Field-Notes 24:727-736.

James, F.C., and N.O. Wamer. ms. Forest bird communities and vegetation structure. (In review).

Johnson, D.H. 1981. How to measure habitat--a statistical perspective. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Techincal Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Johnston, D.W., and E.P. Odum. 1956. Breeding bird populations in relation to plant succession on the Piedmont of Georgia. Ecology 37:50-62.

Karr, J.R. 1968. Habitat and avian diversity on strip-mined land in east-central Illinois. Condor 70:348-357.

Karr, J.R. 1971. Structure of avian communities in selected Panama and Illinois habitats. Ecological Monographs 41:207-233.

Karr, J.R. 1976. Within- and between-habitat avian diversity in African and neotropical lowland habitats. Ecological Monographs 46:457-481.

Karr, J.R., and R.R. Roth. 1971. Vegetation structure and avian diversity in several new world areas. American Naturalist 105:423-435.

Kendeigh, S.C. 1945. Community selection by birds on the Helderberg Plateau of New York. Auk 62:418-436.

Martin, N.D. 1960. An analysis of bird populations in relation to forest succession in Algonquin Provincial Park, Ontario. Ecology 41:126-140.

Niemi, G.J., and L.A Pfannmuller. 1979. Avian communities: approaches to describing their habitat associations. p. 154-178. In DeGraaf, R.M., and K.E. Evans, compilers. Management of north central and northeastern forests for nongame birds: Proceedings of a workshop [Minneapolis, Minn., January 23-25, 1979]. USDA Forest Service General Technical Report NC-51, 268 p. North Central Forest Experiment Station, St. Paul, Minn.

Noon, B.R. 1979. Climax maple-birch-beech forest. American Birds 33:58.

Noon, B.R. 1981. The distribution of an avian guild along a temperate elevational gradient: the importance and expression of competition. Ecological Monographs. 51:105-124.

Noon, B.R. and K.P. Able. 1978. A comparison of avian community structure in the northern and southern Appalachian Mountains. p. 98-117. In DeGraaf, R.M., technical coordinator. Management of southern forests for nongame birds: Proceedings of a workshop [Atlanta, Ga., January 24-26, 1978] USDA Forest Service General Technical Report SE-14, 176p. Southeastern Forest Experiment Station, Asheville, N.C.

Noon, B.R., D.K. Dawson, D.B. Inkley, C.S. Robbins, and S.H. Anderson. 1980. Consistency in habitat preference of forest bird species. Transactions North American Wildlife and Natural Resources Conference 45:226-244.

Oosting, H. 1956. The study of plant communities. 440 p. W. H. Freeman, San Francisco, Calif.

Pfannmuller, L.A. 1979. Bird communities in northeastern Minnesota. M.S. Thesis, 75 p. University of Minnesota, Minneapolis.

Rabe, D. 1977. Structural analysis of woodcocck diurnal habitat in northern Michigan. Proceedings Woodcock Symposium 6:125-134.

Rice, J. 1978. Ecological relationships of two interspecifically territorial vireos. Ecology 59:526-538.

Robbins, C.B. 1978. Determining habitat requirements of nongame species. Transactions North American Wildlife and Natural Resources Conference 43:57-68.

Rotenberry, J.T., and J.A. Wiens. 1978. Nongame bird communities in northwestern rangelands. p. 59-86. In DeGraaf, R.M., technical coordinator. Nongame bird habitat management in the coniferous forests of the western United States: Proceedings of a workshop [Portland, Ore., February 7-9, 1977]. USDA Forest Service General Technical Report PNW-46, 100 p. Pacific Northwest Forest and Range Experiment Station, Portland, Ore.

Roth, R.R. 1976. Spatial heterogeneity and bird species diversity. Ecology 57:773-782.

Schreuder, H.T., and W.L. Hafley. 1977. A useful bivariate distribution for describing stand structure of tree heights and diameters. Biometrics 33:471-478.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58:810-819.

Smith, R.L. 1974. Ecology and field biology. 849 p. Harper and Row, New York, N.Y.

Steel, R.G.D., and J.H. Torrie. 1960. Principles and procedures of statistics. 481 p. McGraw-Hill, New York, NY.

Sturman, W.A. 1968. Description and analysis of breeding habitats of the chickadees, Parus atricapillus and P. rufescens. Ecology 49:418-431.

Thomas, J.W., G.L. Crouch, R.S. Bumstead, and L.D. Bryant. 1975. Silvicultural options and habitat values in coniferous forests. p. 59-86. In Smith, D.R., technical coordinator. Management of forest and range habitats for nongame birds: Proceedings of a symposium [Tuscon, Ariz., May 6-9, 1975]. USDA Forest Service General Technical Report WO-1, 343 p. Washington, D.C.

Thomas, J.W., R.J. Miller, H. Black, J.E. Rodiek, and C. Maser. 1976. Guidelines for maintaining and enhancing wildlife habitat in forest management in the Blue Mountains of Oregon and Washington. Transactions North American Wildlife and Natural Resources Conference 41:452-476.

Titterington, R.W., H.S. Crawford, and B.N. Burgason. 1979. Songbird responses to commerical clear-cutting in Maine spruce-fir forests. Journal of Wildlife Management 43:602-609.

Whitmore, R.C. 1975. Habitat ordination of passerine birds of the Virgin River Valley, southwestern Utah. Wilson Bulletin 87:65-74.

Whitmore, R.C. 1977. Habitat partitioning in a community of passerine birds. Wilson Bulletin 89:253-265.

Whitmore, R.C. 1979. Temporal variation in the selected habitats of a guild of grassland sparrows. Wilson Bulletin 91:592-598.

Wiens, J.A. 1969. An approach to the study of ecological relationships among grassland birds. 93 p. Ornithological Monographs 8.

Wiens, J.A. 1973. Pattern and process in grassland bird communities. Ecological Monographs 43:237-270.

Wiens, J.A. 1974. Habitat heterogeneity and avian community structure in North American grasslands. American Midland Naturalist 91:195-213.

Wight, H.M. 1938. Field and laboratory technique in wildlife management. 105 p. University of Michigan Press, Ann Arbor, Mich.

Appendix I

Life forms and habitat features

A. Life forms and habitat features to be discriminated as line-intercept variables.

Grasses - Narrow-leafed herbaceous plants

Forbs - Broad-leafed herbaceous plants

Woody ground cover - Woody vegetation < 1 m tall

Shrubs - Woody vegetation > 1 m tall and < 3 cm dbh

Saplings - Woody vegetation > 1 m tall and 3 cm ≤ dbh < 8 cm

Trees - Woody vegetation ≥ 8 cm dbh

Litter - Dead plant material excluding downed logs

Water

Bare ground

Rocks

Downed logs - Woody vegetation ≥ 8 cm dbh and ≥ 1.5 m long

B. Tree size classes based on diameter at breast height (dbh) (modified from James and Shugart 1970).

Class Label	Dbh Range (cm)
S	3 ≤ dbh < 8
A	8 ≤ dbh < 15
B	15 ≤ dbh < 23
C	23 ≤ dbh < 38
D	38 ≤ dbh < 53
E	53 ≤ dbh < 69
F	69 ≤ dbh < 84
G	84 ≤ dbh < 102
H	102 ≤ dbh

C. Ground cover life forms to be discriminated in circular forest plots.

Mosses

Ferns

Grasses and sedges - Narrow leafed herbaceous plants

Forbs - Broad leafed herbaceous plants

Woody ground cover - Woody vegetation ≤ 1 m tall

Seedlings - Regeneration from overstory trees, saplings, or shrubs

Litter - Dead plant material excluding slash and logs

Slash and logs - Unrooted woody vegetation (usually dead) lying prostrate

Rocks

Bare ground

Appendix II. Field data sheet for non-forest habitat (≤ 25 percent cover by trees).

Bird sp.	Quad	Plot	Transect	State	County	Latitude ° ' "	Longitude ° ' "	Elevation	Date	Observers

LINE INTERCEPT VARIABLES

TRANSECT INTERVAL	Bare ground	Rocks	Water	Litter	Downed logs	Mosses	Grasses	Forbs	Woody veg. (< 1m tall)	Shrubs (≥ 1m tall, <3cm dbh)	Saplings (3≤dbh<8)	Trees (dbh≥ 8)
1												
2												
3												
4												
5												

NEAREST TREE IN EACH QUARTER

POINT	NE Species	NE Dist.	NE DBH	NE Hgt.	SE Species	SE Dist.	SE DBH	SE Hgt.	SW Species	SW Dist.	SW DBH	SW Hgt.	NW Species	NW Dist.	NW DBH	NW Hgt.
1																
2																
3																
4																
5																

NEAREST SHRUB IN EACH QUARTER

POINT	NE Species	NE Dist.	NE Hgt.	SE Species	SE Dist.	SE Hgt.	SW Species	SW Dist.	SW Hgt.	NW Species	NW Dist.	NW Hgt.
1												
2												
3												
4												
5												

VERTICAL FOLIAGE DENSITY (# of contacts in each height interval)

POINT	0-0.3m N	E	S	W	T	0.3-1 m N	E	S	W	T	1-2 m N	E	S	W	T	2-9 m N	E	S	W	T	> 9 m N	E	S	W	T
1																									
2																									
3																									
4																									
5																									

Appendix III. Field data sheet for forest habitats (> 25 percent cover by trees).

1-4 Bird sp.	5-8 Quad	9-11 Plot	12-15 Circle	16-17 State	18-20 County	Latitude 21-22 ° 23-24 ' 25-26 "	Longitude 27-29 ° 30-31 ' 32-33 "	34-37 Elevation	Distance to edge	42-46 Date	47-52 Observers

1-8 TREE Genus species	9-11 S 3-8 cm	12-14 A 3-15	15-17 B 15-23	18-20 C 23-38	21-23 D 38-53	24-26 E 53-69	27-29 F 69-84	30-32 G 84-102	33-35 H >102

SHRUBS Number of stems intersected

	N 1-7	S 3-4	E 5-6	W 7-8
Decid.				
Conif. 9				

NEAREST TREE & LARGEST LOG IN EACH QUARTER

	Tree Dist. 17-20	Tree DBH 21	Log (≥1.5m long ≥8cm diam.) Dist. 22-25	Log Diam. 26	Log Length 27-29
NE					
SE 30					
SW 43					
NW 56					

DISPERSION INDEX:

Shrubs 69-70 _____ Ground cover 71-72 _____

CANOPY COVER 1-3 [ ] % 4-6 [ ] % coniferous

GROUND COVER 7-9 [ ] %

DENSITY BOARD Count squares at least 50% obscured by green vegetation.

	N 10-11	E 12-13	S 14-15	W 16-17
0-.3m(HH-JJ)				
.3-1m(AA-GG) 18				
1-2m (K-T) 26				
2-3m (A-J) 34				

CANOPY HEIGHT(m)

42-45 Mean	46-49 Max	50-53 Min

SLOPE(°) 54-55 _____

Dominant shrubs(≥ 1m tall, <3cm dbh):

1. 1-8 _____
2. 9 _____
3. 17 _____
4. 25 _____
5. 33 _____

Dominant ground cover (ground-1m) life forms:

1. 41 _____
2. 47 _____
3. 53 _____
4. 59 _____
5. 65 _____

HOW TO MEASURE HABITAT—A STATISTICAL PERSPECTIVE[1]

Douglas H. Johnson[2]

Abstract.--The present workshop reflects the increasing interest in the use of sophisticated statistical analysis for relating wildlife to their habitats. Multivariate methods are useful tools and their results to date have been promising, but closer attention to scientific principles and statistical requirements will prove beneficial, particularly when wildlife-habitat studies are used as the basis for prediction and management of wildlife populations.

This paper discusses several points, some well known in theory but often not fully recognized in practice, others less well known, but ideas that should guide the researcher toward an improved study design that yields more credible results. The major points are 1) The objectives of a study must be clearly thought out and precisely stated in order to design the study properly. 2) Correlation, including its analogs regression and discrimination, is not necessarily causation. 3) If habitat variables are not measured accurately, a host of analytic problems can arise. 4) The reliability and repeatability of habitat measurements are important both statistically and biologically. 5) The question of how many observations are necessary is an open one, but a few methods for addressing the question are offered.

Key words: Correlation; errors in variables; habitat studies; reliability; research design; sample size; scientific method.

INTRODUCTION

The topic I was asked to address can be viewed quite broadly. I chose to focus on some of the problems that have been encountered in my own consultations or have been identified as potentially important in published studies. The following ideas are not new, but, to judge from my experience, they are still worthy of consideration. Nor are the suggestions comprehensive, but attention to them will probably result in improved research designs.

"THE CHOICE OF A SAMPLING STRATEGY DEPENDS ON THE STUDY OBJECTIVES"

This statement is obvious, but, nonetheless, it seems too frequently disregarded, particularly in studies intended by investigators to "learn all we can." I will illustrate the point in a simplified example. Suppose we are interested in bobolinks (Dolichonyx oryzivorus) in grasslands of North Dakota. If I hold as the prime objective an accurate estimate of bobolink breeding density, I might proceed as follows. Stratify the state into regions based on physiography and/or prior

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Statistician, U.S. Fish and Wildlife Service, Jamestown, ND 58401.

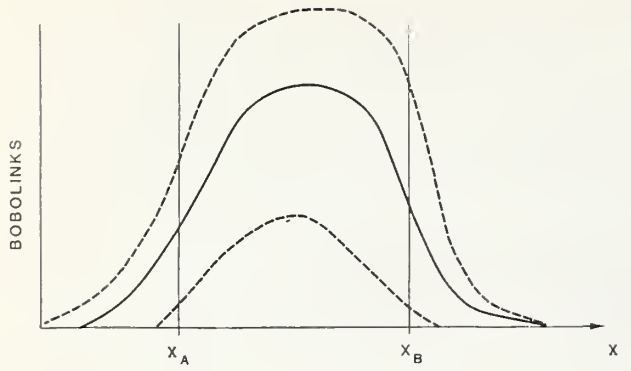


Figure 1. Hypothetical curve indicating the response of a species to an environmental gradient X. Solid line indicates average density; dashed lines denote range in density.

knowledge of bobolink densities. Select sample units of land within each stratum, the sample size proportional to the area of the stratum and to the anticipated standard deviation of bobolink counts. Suppose the bobolink responds to some habitat gradient X, which might be an east-west gradient in precipitation, as shown in figure 1. The average density is indicated by the solid line; the dashed lines denote ranges and reflect variability about the average. To estimate most accurately the mean bobolink density, I would sample most intensively those strata containing land units with values of the habitat gradient between X_A and X_B, where bobolink densities, and their standard deviations, are greatest. I would sample at a low rate the strata of marginal habitats, in which $X < X_A$ or $X > X_B$. Random sampling is necessary here to insure valid estimators of the precision of the mean. As a sideline, I might also measure vegetational features associated with each sample unit.

If, on the other hand, I am not so interested in estimating density as I am in determining relationships between bobolinks and grassland habitat features, perhaps for predictive or management purposes, I would proceed differently. I would then sample the units as much throughout the feasible region as possible, trying to get points all along the gradient. I would be particularly interested in marginal habitats for bobolinks, because some features there are presumably limiting the bird, and those features merit detailed study. Random sampling in this situation may help eliminate misleading results caused by selecting nonrepresentative units.

In general, when the objective of estimating the mean is foremost, we should sample most intensively where birds are common. If we are interested in determining relationships, we should sample more evenly along the gradient of habitat features. This is but a simple example of how specific objectives dictate the design of research and analysis of data that result.

"CORRELATION IS NOT NECESSARILY CAUSATION"

This caveat needs to be remembered as we proceed with linear or nonlinear models involving numerous and often highly intercorrelated variables, all of which are uncontrolled by the investigator. Educational statistics provide a classic example. Among grade school children, performance on scholastic achievement tests is positively correlated with body weight. Yet parents, desirous of improving their childrens' scores, would be ill-advised to fatten them in preparation for testing, because the correlation is spurious: older children tend to perform better in the tests than younger ones, and they also are likely to be heavier. This example may be so obvious as to appear trite, but the analogous situation can occur readily and less overtly in wildlife-habitat studies, in which the plethora of variables and paucity of knowledge about their true relationships can promote misconceptions. Given enough variables and access to a high-speed computer, nearly anyone can find a "significant" association among some of the variables.

"ERRORS IN VARIABLES CAN BE TROUBLESOME"

Consider the linear model

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k + \text{error}. \qquad (1)$$

This form ordinarily represents a regression model but discriminant analysis can be viewed in the same manner if Y is a dummy-variable indicator of group membership (Lachenbruch 1975).

In the usual formulation, the X's are assumed fixed and known quantities, measured without error. In actual practice, however, particularly in ecological work in which the X's are habitat measurements, errors of unknown magnitude often creep in. I personally suspect them to be large more often than not.

Errors in the X's lead to biases in the regression coefficients. A bias is induced whether errors are systematic, reflecting a bias, or simply random, reflecting added variability. Suppose the true relationship between, say, bird density Y and a vector of habitat features $\underline{X}$ is given by equation (1), but the X's are not measured exactly; instead $\underline{Z}$ is measured, where $\underline{Z} = \underline{X} + \underline{\delta}$ and $\underline{\delta}$ is an error of measurement. Many sources of error may contribute to $\underline{\delta}$. For example, sampling variability is usually present; the entire habitat of the animal is not measured, only a sample of it. There is often instrument error; the measured variable may not be exact. Temporal variability may contribute (Whitmore 1979); habitat measurements may not reflect conditions at the time the animals selected the habitat. More generally, the wrong variable may be measured because the correct one is not known.

Suppose that $\underline{Z}$ is unbiased for $\underline{X}$, that is $E(\underline{\delta}) = \underline{0}$, and the variance-covariance matrix of $\underline{\delta}$ is $\underline{D}$, a diagonal matrix. The diagonality implies that errors associated with measurements of different habitat features are uncorrelated.

The regression coefficients are biased (Davies and Hutton 1975, Seber 1977), and the bias vector can be estimated by $n(Z'Z)^{-1}\hat{D}\hat{\beta}$. Z is the matrix containing the $\underline{Z}$ vectors for all observations and n is the sample size. Note that the biases tend to increase in absolute magnitude as $\hat{D}$ becomes large and as Z'Z becomes less well conditioned. That is, the biases increase when measurement errors increase and when highly correlated variables are included. These two conditions are, I believe, very common in ecological practice.

Errors in measured variables affect not only the estimated coefficients, but also their standard errors. The biases here might be either positive or negative, depending on values of the coefficients and the variances (Bloch 1978). Generally, however, the magnitude of the biases will not be unduly large if errors of measurement are reasonably small (Hodges and Moore 1972, Davies and Hutton 1975). Regardless of whether the standard error is biased positively or negatively, in the case of one explanatory variable, the "t" statistic for assessing the significance of that variable will be biased low (Bloch 1978). This feature could cause a truly important habitat variable to be eliminated as nonsignificant. Also, the multiple correlation coefficient may be diminished (Cochran 1970), possibly to the point that the entire set of explanatory variables is nonsignificant.

Errors in measured variables tend to deflate regression coefficients associated with those variables and to lead an ecologist to claim unjustly that their effects are insignificant. The impact of this problem may be particularly severe when the resulting models are used for prediction and/or management. Hodges and Moore (1972) noted that any bias in the regression coefficients will be transmitted into a biased forecast. They also reported several studies in which predictions were available based either on accurately measured predictor variables or on inaccurate (preliminary) values of those variables. The increased error resulting from the use of inaccurate variables was often striking.

The problem of errors in variables is often glossed over by use of a conditional argument. The relationship between Y and $\underline{X}$ is explored, *given* that values of $\underline{X}$ were observed values $\underline{Z}$. This line of reasoning is not clearly stated, perhaps not clearly understood, and certainly of minimal value in prediction and management; the manager is not interested only in habitats possessing those exact values.

The general problem of errors in variables seems to have been ignored in ecology; most applications have been in economics. Moreover, most attention has been given to multiple regression, although limited work has been done with factor analysis (Lawley and Maxwell 1973, Chan 1977).

Davies and Hutton (1975) and Seber (1977:159) presented some working rules for evaluating the effect of errors in variables. If an independent batch of data is available, it is instructive to calculate the *reliability* of each measurement:

$$g_i = \mathrm{Var}(X_i)/\mathrm{Var}(Z_i)$$

$$= \mathrm{Var}(X_i)/[\mathrm{Var}(X_i) + \mathrm{Var}(\delta_i)].$$

Reliability values appreciably below 1 should raise some suspicions about the role of the corresponding variable in the analysis.

At least until the problem is more fully explored, it seems prudent to design ecological studies so that habitat features are accurately measured and relatively independent of one another.

"VARIABLES SHOULD BE MEASURED RELIABLY"

The argument just offered suggests that habitat variables should be measured accurately, in order to reduce D as much as possible and minimize bias in the regression coefficients. There is a further reason for improved reliability: measurements should be repeatable. This feature will become increasingly valuable as studies evolve from their local orientation involving single study areas and are replicated by different researchers in different locations. To truly define the niche of a species requires studies beyond a single woodland; the species must be studied in many parts of its range. To that end, it is mandatory that measurements be reliable and without serious variability due to observer, season, occasion, or other causes.

Precious little is known about intra-observer and inter-observer sources of variability in habitat measurements, and how they compare with differences that are of interest. Ecologists either do not like to make these comparisons or, if they do, prefer not to discuss them.

THE FINAL QUESTION - "HOW LARGE A SAMPLE?"

As a consulting statistician, the first question I am usually asked is how many observations are necessary. I generally respond with one of three numbers: 1, 50, or "great gobs." The former answer, 1, is given occasionally when I believe the hypothesis is stated incorrectly, or in too much generality; a single observation is likely to refute it. The latter answer, great gobs, is given when no hypothesis is at hand, or, if there is one, it is so slippery as to evade capture and possible rejection. Neither answer satisfies the biologist, of course, but either

serves as a necessary prelude to sitting down and thinking about an appropriate hypothesis. The middle answer, 50, or possibly 100 or 10 or 30, is given when the hypothesis is well stated and the experimental procedures thoughtfully described.

I can only give here some general guidelines for determining sample size. First, more observations are needed when the number of variables is large. Many published studies have only slightly more observations than variables, or sometimes even fewer. An appropriate minimum sample size might be 20 observations, plus 3 to 5 more for each variable in the analysis. Larger sample sizes help to overcome difficulties caused by violations of the assumptions of multivariate methods (Green 1979:165).

Second, examine the stability of the estimates, both means and variances. This can be done in several ways, for example, by sampling sequentially, until the mean and variance stabilize. Once the data have been gathered, the stability of the results can be assessed by subsampling, jackknifing, leaving-one-out methods, etc. If the data set is split randomly into two halves, does each half yield conclusions consistent with the other? Better yet, apply the results to another area, or a different year. What predictive value do they have?

Third, investigate the sources of variability, and how they compare in magnitude. Observer variability, temporal variability, variability in measurement method, and others all add up to cloud the variability between features that animals may be responding to. Calls for larger samples are the "knee-jerk" reaction when variability is excessive, but it may be far more advantageous to try instead to reduce this variability by better design.

CONCLUSIONS

It is useful to distinguish two kinds of research, exploratory and confirmatory. In exploratory investigations, the researcher is simply trying to "see what's going on" with respect to a system. This preliminary reconnaissance is for the purpose of hypothesis generation and will likely entail at least a cursory examination of many variables, in order to determine those that might be influential. In exploratory research, a variety of stepwise and ad hoc procedures are acceptable. Results of an exploratory investigation are hypotheses for further testing, not well founded conclusions on which management practices can credibly be based. For the more definitive answers necessary for prediction and management, confirmatory research is needed.

A confirmatory investigation is more in the mold of a classic scientific experiment, in which a hypothesis is stated, an experiment conducted to test that hypothesis, and the outcome used either to reject the hypothesis or to retain it. It is clear that this research, unlike the exploratory investigation, demands a precise hypothesis, clearly stated and unambiguous. The design also needs to be rigorous and the analysis must be statistically correct. A fresh set of data must be brought up; the data used in the exploratory stage to generate a hypothesis cannot be resurrected in the confirmatory stage to test it. Confirmatory research is more difficult to apply to ecological problems than is an exploratory investigation, but management of ecological systems requires the more definitive methods if it is to prove successful. The following suggestions are offered for an investigator planning a confirmatory study of habitat in relation to wildlife. [Green (1979) presented "ten principles" for environmental studies, many of which are equally applicable to wildlife-habitat studies.]

1. Think carefully about the objectives of the study and ask yourself, and other qualified scientists, if the procedures are truly designed to meet those objectives.

2. Remember that correlation (as well as regression and discrimination) is not necessarily causation.

3. Try to obtain habitat measurements that are relatively uncorrelated, to avoid the bias associated with the errors-in-variables problem, and for other good reasons as well. This approach appears far preferable to many ex post facto methods for reducing the number of variables, such as stepwise procedures and principal components analysis.

4. Learn about the kinds of variability in the measurements. Would the same values be obtained tomorrow as today? Would another qualified investigator record the same values? How different would another randomly selected point be? Answers to questions such as these will not only help assess the bias due to the errors-in-variables problem, they will also facilitate cooperative and comparative studies.

5. Sample sizes should be large, especially in relation to the number of variables involved. Samples should be large enough to yield stable and reliable estimates, but methods of reducing the inherent variability of measurements may be more fruitful than simply increasing the sample size.

6. And finally, for the ecologist, consult your friendly neighborhood statistician. Consult him early, in the project design phase. Consult him often, as data-gathering proceeds. Then, when you consult with him about analysis and interpretation, he will be of much better humor and of far greater benefit to you.

ACKNOWLEDGMENTS

I am grateful to Drs. L.M. Cowardin and S.G. Machado for comments on an earlier draft of this report.

LITERATURE CITED

Bloch, F.E. 1978. Measurement error and statistical significance of an independent variable. American Statistician 32(1):26-27.

Chan, N.N. 1977. On an unbiased predictor in factor analysis. Biometrika 64:642-644.

Cochran, W.G. 1970. Some effects of errors of measurement on multiple correlation. Journal of the American Statistical Association 65: 22-34.

Davies, R.B., and B. Hutton. 1975. The effect of errors in the independent variables in linear regression. Biometrika 62:383-391.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. 257 p. John Wiley and Sons, New York, N.Y.

Hodges, S.D., and P.G. Moore. 1972. Data uncertainties and least squares regression. Applied Statistics 21:185-195.

Lachenbruch, P.A. 1975. Discriminant analysis. 128 p. Hafner Press, New York.

Lawley, D.N., and A.E. Maxwell. 1973. Regression and factor analysis. Biometrika 60:331-338.

Seber, G.A.F. 1977. Linear regression analysis. 465 p. John Wiley and Sons, New York, N.Y.

Whitmore, R.C. 1979. Temporal variation in the selected habitats of a guild of grassland sparrows. Wilson Bulletin 91:592-598.

DISCUSSION

JIM WOEHR: Is a distribution of plots systematically with respect to space a random sample of plants?

DOUGLAS JOHNSON: No, but it may be adequately represented by a model of randomness. The crucial issue is to define the population of which the sample is representative. Random sampling insures that condition if the sample is infinitely large, although finite ones can certainly be misrepresentative. The Bayesian concept of exhangeability is analogous.

The immediate question is whether the plants might conceivably vary sysmatically in space. In North Dakota, for example, plots located 1 mile apart might not give a representative portrayal of the plants. If the initial transect was along a roadside, most subsequent ones would be also, and smooth brome (Bromis inermis), for instance, would appear much more commonly in the plots than in the state as a whole. An interval of a different length between plots could yield rather accurate results, however.

Multivariate Methods

DISCRIMINANT ANALYSIS IN WILDLIFE RESEARCH: THEORY AND APPLICATIONS[1]

Byron Kenneth Williams[2]

Abstract.--Discriminant analysis, a method of analyzing grouped multivariate data, is often used in ecological investigations. It has both a predictive and an explanatory function, the former aiming at classification of individuals of unknown group membership. The goal of the latter function is to exhibit group separation by means of linear transforms, and the corresponding method is called canonical analysis. This discussion focuses on the application of canonical analysis in ecology. In order to clarify its meaning, a parametric approach is taken instead of the usual data-based formulation. For certain assumptions the data-based canonical variates are shown to result from maximum likelihood estimation, thus insuring consistency and asymptotic efficiency.

The distorting effects of covariance heterogeneity are examined, as are certain difficulties which arise in interpreting the canonical functions. A "distortion metric" is defined, by means of which distortions resulting from the canonical transformation can be assessed. Several sampling problems which arise in ecological applications are considered. It is concluded that the method may prove valuable for data exploration, but is of limited value as an inferential procedure.

Key words: Canonical analysis; covariance heterogeneity; discriminant analysis; eigenvector.

INTRODUCTION

Discriminant analysis is a technique which has come to be much used in ecological investigations. It is applicable to the study of niche breadth, niche overlap, resource partitioning, habitat selection, community structure, and many other topics. In fact the methodology is potentially useful for any ecological situation in which an association is desired between well defined groups and a set of ecologically meaningful measurements.

The data for a discriminant analysis comes to the investigator in the form of a categorical "response" variate and a corresponding set of (usually continuous) "predictor" variates. One of the objectives of the analysis is to predict the category to which an observation belongs, based on values of the predictor variates and an appropriate underlying statistical model. Such a formulation is essentially classificatory, and the prediction equations which result are called classification functions. Alternatively, the

[1]Paper presented at The Use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Biometrician, U.S. Fish and Wildlife Service, Migratory Bird and Habitat Research Laboratory, Laurel, MD 20811.

Table 1. Examples of discriminant analysis from the ecological literature. A number of measurements are made on each sample, and samples are aggregated into groups according to a grouping index.

Groups defined by	Observation measurements	Authors
Faunal species	habitat structural characteristics	Green (1971, 1974); James (1971); Bertin (1977); Cody (1978); Dueser and Shugart (1978, 1979)
Vegetation species	faunal species abundances	Ricklefs (1977)
Species presence/ absence	habitat structural characteristics	Anderson and Shugart (1974); Conner and Adkisson (1976)
Animal behavior	habitat structure, climate	Conroy et al. (1979)
Season	photosynthetic rates	Kowal et al. (1972)
Species and sex	behavioral measurements	Conley (1976)
Geographic area	vegetation densities	Norris and Barkham (1970)
Abiotic categories	habitat factors	Smith (1977)
Artificial classes	vegetation densities	Grigal and Goldstein (1971); Goldstein and Grigal (1972)
Faunal species	meristic and morphometric characteristics	Montanucci (1978)
Faunal species	song and feeding behaviors 1978c)	Rice (1978a, 1978b,
Geographic area	faunal abundance	Buzas (1967)
Soil groups	chemical concentrations	Horton et al. (1978)
Socially defined breeding demes	body measurements	Buechner and Roth (1974)

objective of the discriminant analysis may be to establish optimal "separation" of groups, based on certain linear transforms of the predictor variates. This latter approach aims at interpretation as well as prediction, and the linear functions used to explain group separation are called canonical variates. As indicated below, under certain distribution assumptions the classification approach to discriminant analysis is logically consequent to the canonical analysis. It can be shown (Williams, publication submitted) that the canonical variates themselves may be used to develop a classification procedure entirely equivalent to that produced by the predictive methodology.

Applications of discriminant analysis in the ecological literature are many and varied. A substantial proportion of these, though by no means all, concern the assessment of species-habitat associations. There is a preponderance of applications to avifauna and, to a lesser degree, to small mammals. Many are single-species studies in which groups are determined by presence or absence of an individual species. Table 1 displays some examples of the use of discriminant analysis reported in the literature. As indicated in the table, the grouping index can range over many different attributes, from vegetation types to faunal species to artificial classes built by clustering procedures. The corresponding measurement variables can range over a variety of habitat measurements, such as plant densities or vegetation structure characters, and faunal measurements such as species abundance or behavioral and morphological characteristics.

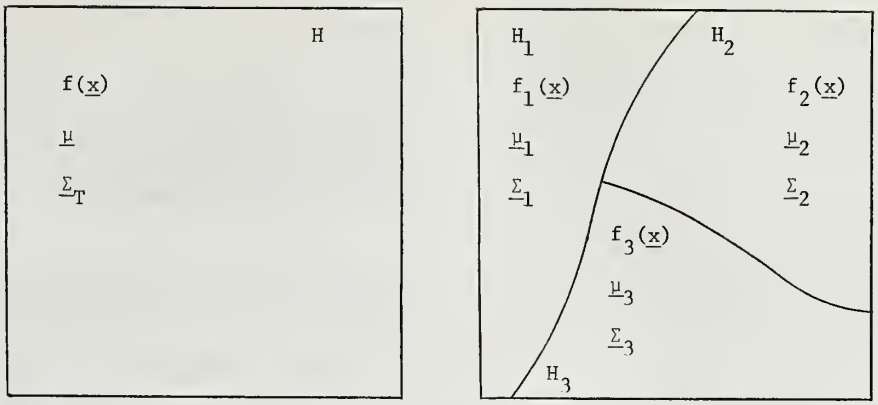


Figure 1. Partition of an environment by species utilization. (a) The unpartitioned environment, at each point of which there corresponds a vector $\underline{x}$. (b) The environment partitioned into species-specific habitats.

Most of these studies involve the use of canonical variates at some point in the investigation. Observations are plotted in a "canonical space" of reduced dimensionality, and the patterns displayed are interpreted. Attempts generally are made to interpret the canonical functions themselves, often by an examination of coefficients. These efforts are hampered by certain statistical and conceptual difficulties which arise in applications with ecological data.

In view of the increasing use of canonical variates in ecological investigations, this paper focuses on the theory and applications of canonical analysis. For purposes of elucidation the general probability model for discriminant analysis is introduced below, along with a brief description of linear classification based on it. Then canonical analysis is investigated from the point of view of vector transformations. My goal is to specify precisely the model, the estimation procedures and the geometric logic of canonical analysis. A geometric approach, by clarifying the structure of canonical analysis, reveals the importance of its statistical assumptions and the possible consequence of their violation.

PROBABILITY STRUCTURE

The probability structure to which discriminant analysis applies is specified as follows. Consider a mixture of g populations $\pi_1,\ldots,\pi_g$ with mixing proportions $q_1,\ldots q_g$, represented by the vector $\underline{q}$. Associated with each individual in the mixture is a vector $[i,\underline{x}']'$ of p+1 components, the first of which (i) specifies population membership. The remaining p components in $\underline{x}$ are a set of measurement values associated with the individual. Random selection from the mixture defines a random variable I for population membership and a random vector $\underline{X}$ of measurement values. The probability structure of the mixture is defined by a set of distributions:

(i) $\quad P(I = i) = q_i$

the probability that an individual chosen at random from the mixture is in population π_i (the values in $\underline{q}$ are often called prior probabilities);

(ii) $\quad f_{X|I}(\underline{x}|i) = f_i(\underline{x})$,

the conditional distribution of $\underline{X}$ over the population π_i. If the set of conditional distributions $f_1(\underline{x}),\ldots,f_g(\underline{x})$ is given as the vector $\underline{f}$, then the mixture is represented by an ordered pair $(\underline{f},\underline{q})$.

For the ensuing discussion it may be of value to think of the sampling universe, denoted by H, as a habitat which is partitioned by way of species utilization. A set of habitat variables is measured on each sample plot, for example canopy height, ground cover and shrub density. These variables have a frequency distribution $f(\underline{x})$ over the habitat H with mean $\underline{\mu}$ and covariance $\underline{\Sigma}_T$.

In addition, each plot has associated with it a variable which indicates which species utilizes it. This variable partitions H into specific habitats, each with its own frequency distribution $f_i(\underline{x})$ of habitat variables. The proportion of H which is included in the habitat of species i is q_i. Such a partitioning is shown in figure 1 for three groups.

Figure 1(a) indicates the sampling universe H, over which can be measured vector values of $\underline{x}$. The distribution of $\underline{x}$ over H is given by $f(\underline{x})$, with mean $\underline{\mu}$ and covariance $\underline{\Sigma}_T$. No partitioning principle is involved in this distribution. In Figure 1(b) H has been partitioned into three subsets. The values that $\underline{x}$ can take in subset H_i define the conditional distribution $f_i(\underline{x})$ with mean $\underline{\mu}_i$ and covariance $\underline{\Sigma}_i$. Mixing proportions q_i specify the proportion of H constituted of H_i, and the relationship of moments in partitioned and unpartitioned spaces is given by

$$\underline{\mu} = q_1\underline{\mu}_1 + \ldots + q_g\underline{\mu}_g$$

and

$$\underline{\Sigma}_T = \overline{\underline{\Sigma}} + \underline{A},$$

where

$$\overline{\underline{\Sigma}} = q_1\underline{\Sigma}_1 + \ldots + q_g\underline{\Sigma}_g$$

and

$$\underline{A} = \sum_{i=1}^{g} q_i(\underline{\mu}_i - \underline{\mu})(\underline{\mu}_i - \underline{\mu})'.$$

It is, of course, not required of the specific habitats that their corresponding measurements $\underline{x}$ be grouped into disjoint sets. H_i and H_j may both contain plots characterized by the same measurements $\underline{x}$, and in fact the range of measurements may be identical over H_i and H_j. It is necessary, however, that the statistical distribution of habitat values differs across species. Discriminant analysis seeks to highlight these among-group statistical differences.

CLASSIFICATION PROCEDURES

The classification problem may be stated in terms of the structure exhibited above: to predict population membership for an individual chosen from $(\underline{f},\underline{q})$, given that $\underline{X} = \underline{x}$. Stated differently, the problem is to classify individuals into one of the g populations based on observed values $\underline{x}$. Classification is determined by

$$P[I = i|\underline{X} = \underline{x}],$$

the probability that a sampling unit with observed values $\underline{x}$ is a member of population π_i. A procedure which minimizes classification error is to assign an individual to π_i if

$$P[I = i|\underline{X} = \underline{x}] = \max \{P[I = j|\underline{X} = \underline{x}]: j = 1,\ldots,g\}.$$

Since

$$P[I=i|\underline{X} = \underline{x}] = \frac{q_i f_i(\underline{x})}{\sum_{j=1}^{g} q_j f_j(\underline{x})}$$

this procedure is equivalent to classification based on

$$q_i f_i(\underline{x}) = \max \{q_j f_j(\underline{x}): j=1,\ldots,g\}.$$

The earliest and best developed discriminant methodology assumes a multivariate normal distribution. That is, the conditional distribution of predictor variates is given by

$$f_i(\underline{x}) = (2\pi)^{-p/2} |\underline{\Sigma}_i|^{-1/2} \exp[-1/2(\underline{x}-\underline{\mu}_i)'\underline{\Sigma}_i^{-1}(\underline{x}-\underline{\mu}_i)],$$

where i indexes population π_i and $\underline{\mu}_i$, $\underline{\Sigma}_i$ are the corresponding mean vector and covariance matrix. Linear discrimination results for the assumption of covariance homogeneity:

$$\underline{\Sigma}_i = \underline{\Sigma}, \quad i=1,\ldots,q.$$

In this case the conditional distributions can be rewritten in group-specific terms:

$$f_i(\underline{x}) = (2\pi)^{-p/2}|\underline{\Sigma}|^{-1/2}\exp[1/2\underline{x}'\underline{\Sigma}^{-1}\underline{x}] \cdot \exp[(\underline{x} - 1/2\ \underline{\mu}_i)'\underline{\Sigma}^{-1}\underline{\mu}_i]$$

$$= w(\underline{x})\exp[(\underline{x} - 1/2\ \underline{\mu}_i)'\underline{\Sigma}^{-1}\underline{\mu}_i],$$

where $w(\underline{x})$ is free of group-specific parameters $\underline{\mu}_i$.

For these distributions the optimal classification rule may be simplified by noting that

$$q_i f_i(\underline{x}) \geq q_j f_j(\underline{x})$$

if and only if

$$\ln q_i + \ln f_i(\underline{x}) \geq \ln q_j + \ln f_j(\underline{x}).$$

Since

$$\ln f_i(\underline{x}) = \ln w(\underline{x}) + (\underline{x} - 1/2\ \underline{\mu}_i)'\underline{\Sigma}^{-1}\underline{\mu}$$

and $\ln w(\underline{x})$ has no discriminating power, the classification rule can be specified with the linear function

$$L_i(\underline{x}) = (\underline{x} - 1/2\ \underline{\mu}_i)'\underline{\Sigma}^{-1}\underline{\mu}_i + \ln q_i.$$

The procedure is to classify an observation into group i, based on the observed values $\underline{x}$, if

$$L_i(\underline{x}) = \max \{L_j(\underline{x})|j=1,\ldots,g\}.$$

This is simply Fisher's linear discriminant function (Fisher 1936). When $\underline{\mu}_i$, $\underline{\Sigma}$, and q_i are unknown, the maximum likelihood estimates may be used in the expression for $L_i(\underline{x})$. A substantial body of literature exists on the statistical consequences, especially in error rate analysis, of this replacement (see, e.g., Toussaint 1974).

SAMPLE-BASED CANONICAL ANALYSIS

A different approach to discrimination, which sometimes yields additional information about group differences, is based on the use of canonical transformations. The usual data-based methods of canonical analysis involve the determination of linear transforms

$$z = \underline{a}'\underline{x}$$

which maximize

$$\sum_{i=1}^{g} n_i(\bar{z}_i - \bar{\bar{z}})^2 / \sum_{i=1}^{g}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2,$$

where

$$\bar{z}_i = \underline{a}'\bar{\underline{x}}_i$$

and

$$\bar{\bar{z}} = \underline{a}'(1/n \sum_{i=1}^{g} n_i\bar{\underline{x}}_i).$$

In these expressions n_i is the sample size for population π_i and $n = \sum_{i=1}^{g} n_i$. If the data for a sample from the mixture $(\underline{f},\underline{q})$ are aggregated by

$$\underline{T} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underline{x}_{ij} - \bar{\underline{x}})(\underline{x}_{ij} - \bar{\underline{x}})'$$

$$= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)'$$

$$+ \sum_{i=1}^{g} n_i(\bar{\underline{x}}_i - \bar{\underline{x}})(\bar{\underline{x}}_i - \bar{\underline{x}})'$$

$$= \underline{W} + \underline{B},$$

then the procedure above is formally equivalent to finding eigenvectors corresponding to the g-1 non-zero eigenvalues from

$$[\underline{B} - \lambda\,\underline{W}]\,\underline{a} = \underline{0}.$$

These eigenvectors define g-1 canonical variates

$$z_i = \underline{a}_i'\underline{x},\ i=1,\ldots,g-1$$

which are often described as linear transforms which "maximize among-group variation relative to within-group variation."

PARAMETRIC CANONICAL ANALYSIS

Under certain probability conditions this intuitive and essentially non-parametric method can be generated by a parametric procedure which considers linear transformations of population means. The goal once again is to establish separation of group means, but in this instance the focus initially is on population differences and the parameters characterizing them. Data enters the analysis by way of parameter estimation, at which point the appropriate computing forms can be developed. This approach in some sense reverses the foregoing data based analysis: rather than manipulating data into a statistic by which to interpret sample structure, a geometric focus emphasizes the parametric structure which then is estimated with the data. The intended result from this orientation is a sharpening of the assumptions for canonical analysis and a clarification of its utility for inferences in ecological studies.

Consider the group means $\underline{\mu}_i$, i=1,...,g and common dispersion $\underline{\Sigma}$. We seek in the canonical analysis to describe the separation of these means in a statistically meaningful way. The global mean of the population mixture $(\underline{f}, \underline{q})$ is given by

$$\underline{\mu} = \sum_{i=1}^{g} q_i\underline{\mu}_i,$$

and the population deviations $\underline{\mu}_i - \underline{\mu}$ are defined by the difference between group and overall means. Let L be any arbitrary line through $\underline{\mu}$ with direction cosines $\underline{a}$. Then the projection of $\underline{\mu}_i - \underline{\mu}$ onto L is a vector with squared length given by

$$d_i^2 = [\underline{a}'(\underline{\mu}_i - \underline{\mu})]^2$$

$$= \underline{a}'(\underline{\mu}_i - \underline{\mu})(\underline{\mu}_i - \underline{\mu})'\underline{a}.$$

The average of squared lengths is

$$\sum_{i=1}^{g} q_i d_i^2 = \sum_{i=1}^{g} q_i\underline{a}'(\underline{\mu} - \underline{\mu})(\underline{\mu} - \underline{\mu})'\underline{a}$$

$$= \underline{a}'\left[\sum_{i=1}^{g} q_i(\underline{\mu}_i - \underline{\mu})(\underline{\mu} - \underline{\mu})'\right]\underline{a}$$

$$= \underline{a}'\underline{A}\,\underline{a},$$

which is maximum for some particular direction $\underline{a}^*$ of the projection line. It can be shown that $\underline{a}^*$ is the dominant eigenvector of $\underline{A}$. The optimal direction which is orthogonal to $\underline{a}^*$ is given by the second dominant eigenvector of $\underline{A}$, and so on. Since $\underline{A}$ is of rank g-1 there exist g-1 orthogonal directions by which to separate means, corresponding to the non-zero eigenvalues of $\underline{A}$.

The use of $\underline{a}^*$ (and the remaining eigenvectors) to separate group means is optimal in the sense that for no other projection line is the average of squared distances between means as large. This method is suboptimal, however, in that it does not account for variances and covariances within each population. The failure to accomodate the covariance structure can lead to two highly undesirable results. First, variates with high variation (and therefore low information content) have the same influence on the analysis as do variates with low variation (and high information content). Second, the effect of weighting highly correlated variables equally is to base the analysis less on statistical content than merely on the number of variates included in it. In such a situation one could force group separations to reflect any arbitrarily chosen variate merely by including additional positively

correlated variates in the analysis. Such distortions and abritrariness are precisely the kinds of effects which motivate the use of $\underline{\Sigma}^{-1}$ in the Mahalanobis distance formula.

Canonical analysis includes the same adjustment, the effect of which is to eliminate covariances and unequal variances. This adjustment utilizes a square-root factorization $\underline{\Sigma}^{1/2}$ of $\underline{\Sigma}$ to define new variables

$$\underline{x}^* = \underline{\Sigma}^{-1/2}\underline{x}$$

with conditional mean $\underline{\Sigma}^{-1/2}\underline{\mu}_i$ and identity covariance (Graybill 1976). Thus the original variates are transformed into unit variance, uncorrelated variates, eliminating ambiguities and distortions.

The same projection argument as above may be used on the transformed means $\underline{\Sigma}^{-1/2}\underline{\mu}_i$. The corresponding deviations are

$$\underline{\Sigma}^{-1/2}\ (\underline{\mu}_i - \underline{\mu}),$$

and the least squares lines fitting these deviations are based on a spectral decomposition of

$$\sum_{i=1}^{g} q_i\ \underline{\Sigma}^{-1/2}\ (\underline{\mu}_i - \underline{\mu})(\underline{\mu}_i - \underline{\mu})'\ \underline{\Sigma}^{-1/2}$$

$$= \underline{\Sigma}^{-1/2}\ \underline{A}\ \underline{\Sigma}^{-1/2'}.$$

The corresponding matrix equation is

$$[\underline{\Sigma}^{-1/2}\ \underline{A}\ \underline{\Sigma}^{-1/2'} - \lambda\ \underline{I}]\ \underline{v} = \underline{0},$$

which may be written as

$$\underline{\Sigma}^{-1/2}\ [\underline{A} - \lambda\ \underline{\Sigma}]\ \underline{\Sigma}^{-1/2'}\ \underline{v} = \underline{0}$$

or

$$[\underline{A} - \lambda\ \underline{\Sigma}]\ \underline{u} = \underline{0}, \qquad (1)$$

where

$$\underline{u} = \underline{\Sigma}^{-1/2'}\underline{v}.$$

Canonical variates are then given by

$$\underline{z} = \underline{v}'\underline{x}^*$$
$$= \underline{v}'\ \underline{\Sigma}^{-1/2}\ \underline{x}$$
$$= \underline{u}'\ \underline{x}.$$

There are g-1 such variates, corresponding to the non-zero eigenvalues of equation (1). They may be represented by the canonical transform

$$\underline{z} = \underline{U}'\underline{x}, \qquad (2)$$

where the columns of $\underline{U}$ are vector solutions of equation (1).

In most situations none of the parameters for the population mixture is known with certainty, and they must be estimated from the data. Assume that a random sample of size n from the population mixture is obtained, n_i of which are from population π_i. Based on the multivariate normal assumption the following maximum likelihood estimates result:

$$\hat{q}_i : n_i/n$$

$$\hat{\underline{\mu}}_i : \bar{\underline{x}}_i = 1/n_i \sum_{j=1}^{n_i} \underline{x}_{ij}$$

$$\hat{\underline{\Sigma}} : \underline{S} = 1/(n-g) \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)'$$

$$\hat{\underline{\mu}} : \sum_{i=1}^{g} \hat{q}_i\ \hat{\underline{\mu}}_i = 1/n \sum_{i=1}^{g} \sum_{j=1}^{n_i} \underline{x}_{ij}$$

$$\hat{\underline{A}} : 1/n\ \underline{B} = 1/n \sum_{i=1}^{g} n_i\ (\bar{\underline{x}}_i - \bar{\underline{x}})(\bar{\underline{x}}_i - \bar{\underline{x}})' \quad .$$

When these maximum likelihood estimates are used in equation (1) in place of the population parameters, the equation becomes

$$[\hat{\underline{A}} - \lambda\ \hat{\underline{\Sigma}}]\ \underline{u} = \underline{0} \qquad (3)$$

or

$$[1/n\ \underline{B} - \lambda\ \underline{S}]\ \underline{u}$$

$$= 1/n\ [\underline{B} - n/(n-g)\ \lambda\ \underline{W}]\ \underline{u}$$

$$= \underline{0}\ .$$

Equation (3) has the same eigenstructure as does

$$[\underline{B} - \lambda\ \underline{W}]\ \underline{u} = \underline{0}.$$

The effect of n/(n-g) is to scale the eigenvalues, leaving eigenvectors unchanged. This can be seen by

$$[\underline{B} - \lambda\ n/(n-g)\ \underline{W}\ ]\ \underline{u}$$
$$= \underline{W}^{1/2}\ [\underline{W}^{-1/2}\ \underline{B}\ \underline{W}^{-1/2'} - \lambda\ n/(n-g)\ \underline{I}]\ \underline{W}^{1/2'}\underline{u},$$

from which it follows that n/(n-g) λ is an eigenvalue and $\underline{W}^{1/2'}\underline{u}$ the corresponding eigenvector of $\underline{W}^{-1/2}\ \underline{B}\ \underline{W}^{-1/2'}$. Therefore eigenvectors of

$$[\hat{\underline{A}} - \lambda\ \hat{\underline{\Sigma}}]\ \underline{u} = \underline{0}$$

and

$$[\underline{B} - \lambda\ \underline{W}]\ \underline{u} = \underline{0}$$

are identical. This demonstrates the equivalence

of parametric and data-based approaches to canonical analysis, given the following three assumptions:

(i) the data consist of a random sample from a population mixture of multivariate normal populations;
(ii) covariances are homogeneous across all populations;
(iii) maximum likelihood estimates are used in place of parametric values.

In ecological investigations it is hoped that an intelligent sampling plan can be combined with the central limit theorem to approximately satisfy condition (i). Also, the many advantages of maximum likelihood estimation make condition (iii) a reasonable procedure. Therefore the applicability of canonical analysis in ecological investigations seems to hinge on condition (ii), the homogeneity of covariance.

It is noted parenthetically that the canonical variates in equation (2) are uncorrelated and have a variance of unity. Furthermore, it can be shown that

$$\underline{u}_i' \underline{A} \underline{u}_i = \lambda_i, \quad i=1,\ldots,g-1$$

(Williams, publication submitted). Two extremely useful invariance properties follow (let the transformed population means be represented by $\underline{n}_i = \underline{U}'\underline{\mu}_i$, i=1...,g):

First, relative distances are maintained in canonical space. That is, if Mahalanobis distances in observation and canonical space are defined by

$$D_i(\underline{x}) = (\underline{x} - \underline{\mu}_i)' \underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_i)$$

and

$$D_i(\underline{z}) = (\underline{z} - \underline{n}_i)'(\underline{z} - \underline{n}_i)$$

respectively, then

$$D_i(\underline{x}) - D_j(\underline{x}) = D_i(\underline{z}) - D_j(\underline{z})$$

(Williams, publication submitted). Second, group mean differences are also maintained:

$$(\underline{\mu}_i - \underline{\mu}_j)' \underline{\Sigma}^{-1} (\underline{\mu}_i - \underline{\mu}_j)$$

$$= (n_i - n_j)'(n_i - n_j).$$

This last result is of obvious importance in the study of resource partitioning and niche overlap (MacArthur and Levins 1967, Harner and Whitmore 1977).

INEQUALITY OF COVARIANCES

It remains to rationalize the procedures of canonical analysis under the assumption of unequal conditional covariance matrices. Given conditions (i) - (iii) above, the canonical analysis represents a statistically meaningful attempt to optimally separate population means. Its meaning when condition (ii) is violated becomes more problematic, and obviously depends on what matrix $\underline{\Sigma}$ is used in equation (1). It should be noted that the square root transform, invoked to eliminate covariances, will most certainly fail to have its intended effect when group-specific covariances are unequal. Indeed, no matter what "covariance" matrix $\underline{\Sigma}$ is used in the analysis, the resulting conditional covariances are still heterogeneous and have the form

$$\underline{\Sigma}^{-1/2} \underline{\Sigma}_i \underline{\Sigma}^{-1/2'} .$$

In effect nothing of value for covariance stabilization has been gained by the transform, and the meaningfulness of the canonical procedure is thrown into question.

An often used approach is simply to use the overall mixture covariance $\underline{\Sigma}_T$, or to use an "average" covariance defined by

$$\underline{\Sigma} = q_1 \underline{\Sigma}_1 + \ldots + q_g \underline{\Sigma}_g. \qquad (4)$$

That these two matrices yield the same eigenstructure can be easily shown:

Since

$$\underline{\Sigma}_T = \overline{\underline{\Sigma}} + \underline{A}$$

we have

$$[\underline{A} - \lambda \underline{\Sigma}_T] \underline{u}$$

$$= [\underline{A} - \lambda (\overline{\underline{\Sigma}} + A)] \underline{u}$$

$$= [(1 - \lambda) \underline{A} - \lambda \overline{\underline{\Sigma}}] \underline{u}$$

$$= (1 - \lambda) [\underline{A} - \lambda/(1 - \lambda) \overline{\underline{\Sigma}}] \underline{u}$$

$$= \underline{0}.$$

Therefore $\underline{\Sigma}_T$ and $\overline{\underline{\Sigma}}$ yield the same eigenvectors when used in equation (1). Note that when conditional covariances are identical, $\underline{\Sigma}$ in equation (4) is simply their common value, and the equation

$$[\underline{A} - \lambda \underline{\Sigma}_T] \underline{u} = \underline{0}$$

is equivalent in its eigenvectors to equation (1). When covariances are unequal, however, both $\overline{\underline{\Sigma}}$ and $\underline{\Sigma}_T$ are defined only over the population mixture, and neither generates an eigenstructure equivalent to that based on individual dispersions $\underline{\Sigma}_i$.

One result of heterogeneous covariances is that the canonical variates are uncorrelated over the mixture ($\underline{f}$, $\underline{g}$), but not over the conditional distributions. For any given conditional population they may be highly correlated, and this greatly complicates their interpretation. Furthermore, the representation of conditional

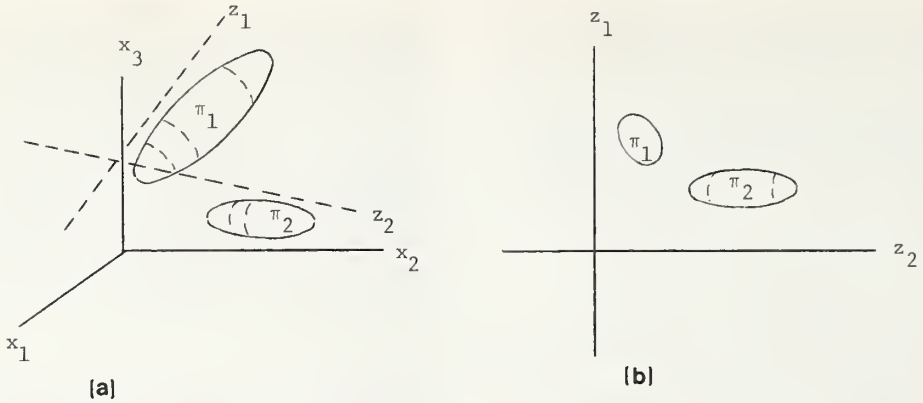


Figure 2. Population representatives in observation and canonical space. (a) The dispersion of two populations in observation space. Canonical axis are superimposed by dotted lines. (b) The sample population dispersions in canonical space.

populations in the g-1 dimensional space defined by a canonical analysis may severely distort their geometric configurations. An example is shown in figure 2.

Figure 2(a) shows two conditional distributions in observation space, one (π_1) with its dominant axis nearly orthongonal to the canonical axes z_1 and z_2. Representation of these populations in canonical space is shown in figure 2(b). Such group specific distortions result from the projection of p-variate distributions into g-1 dimensions, and they greatly complicate the interpretation of areas in canonical space. It follows that one must exercise caution in analyzing conditional population dispersions with canonical variates. This is not unexpected, since the fundamental purpose of discrimination is the assessment of group differences, rather than the analysis of dispersion.

INTERPRETATION OF CANONICAL VARIATES

The canonical analysis assumes its most obvious value for ecologists by providing interpretable transforms of data. As mentioned above, these transforms are specifically chosen to separate group means in an optimal manner. The canonical variates which result are uncorrelated linear combinations of observation variables, the coefficients in which may be given ecological interpretations. At this point in the analysis biological insight is of fundamental importance. It is, however, important to recognize certain limitations on interpretability with these transforms.

Both magnitude and sign of the canonical coefficients often are used for interpretation. Under certain conditions, e.g., when one coefficient dominates all others in the canonical transform and the correlation structure is simple, this is an acceptable practice. In such a situation it is fairly straightforward to assess the "meaning" of the canonical variate, i.e., to specify what it represents biologically. When the correlation structure is complex and there are several coefficients of significant size, however, interpretation is not so direct. The difficulty arises from the fact that the observation variables in the canonical transforms are correlated, some of them perhaps highly correlated. Individual canonical coefficients in this case reflect not only the influence of their corresponding observation variables, but also the influence of other variables as reflected through the correlation structure of the data. Most people who have worked with discriminant analysis have probably seen cases in which positively correlated variates have canonical coefficients with different signs. These apparently inconsistent coefficients indicate that the two corresponding variables include information from the remaining observation variates which influences the canonical variate in opposing ways. It can be shown (Williams, in prep.) that the correlation of the canonical variate z_i and an observation variate x_j is

$$\text{corr}\,(z_i, x_j) = 1/\sqrt{\lambda_i}\,\left(a^*_{ij} + \sum_{k \neq j} a^*_{ik}\,\rho_{kj}\right), \quad (5)$$

where a^*_{ij} is the "standardized" canonical coefficient of x_j in the i^{th} canonical transform. This expression reveals that the effect of an observation variable on the canonical variate is only partially given by the numerical value of its corresponding coefficient. Terms involving the remaining coefficients and the correlation structure between variables also influence the association between z_i and x_j, and sometimes this latter influence is predominant. It follows that one cannot safely interpret the coefficients singly. A similar argument can be made against interpreting pairs, triples, or any subset of

coefficients in a canoncal function. As one might expect, this same problem also arises in ordinary least squares regression: the values of individual parameters reflect the correlation structure of the data. For complex structures, magnitudes and even signs of coefficients are dependent on what additional variables are included in the model (Weiner and Dunn 1966). It makes little sense in that case to base one's interpretations on individual coefficients. A safer technique is to examine the correlation of the canonical variate either with individual observation variables included in the canonical analysis (5), or with ancillary information not included in it (Green 1971, James 1971, Dueser and Shugart 1978). High correlations may then provide an interpretable "meaning" of the canonical variate.

There is in the literature one other method of interpretation, about which some words of caution should be voiced. This involves the use in canonical space of "equal frequency ellipses", defined in multivariate populations by hyperellipses of constant probability density (Harner and Whitmore 1977). For a given distribution with moments ($\underline{\mu}_i$, $\underline{\Sigma}_i$), a point in observation space will be on only one such ellipse. When g populations are defined in a discrimination analysis, the observation lies on ellipses which are specific to each population. It is noted that when populations are normally distributed these ellipses are defined by Mahalanobis distances. They may be used to generate a classification methodology identical with the usual classification procedure as outlined above.

Several results concerning equal frequency ellipses follow from the equal covariance assumption. First, they intersect along straight lines in the observation space. Second, the optimal classification procedure based on them is linear. These two properties are of course equivalent, since the linear classification functions correspond to a linear partitioning of the observation space. Third, the relative distances of points in canonical space as measured by $D_i(\underline{z})$, are identical to those in observation space (Williams, publication submitted). This effectively means that equal frequency ellipses in canonical space can be used to assess the position of an observation relative to the group means: there is, in essence, no important "loss of information" about conditional distributions when canonical variates are used. Thus ecologists are justified in using the canonical variates rather than the original data to investigate ecological distributions. The invariance property is indicated below with Mahalanobis distances, using three group means and an observation vector $\underline{x}$:

Ranking	Observation Space	Canonical Space
1	$(\underline{x}-\underline{\mu}_i)'\underline{\Sigma}^{-1}(\underline{x}-\underline{\mu}_1)$	$(\underline{z}-\underline{\eta}_1)'\underline{V}^{-1}(\underline{z}-\underline{\eta}_1)$
2	$(\underline{x}-\underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{x}-\underline{\mu}_2)$	$(\underline{z}-\underline{\eta}_2)'\underline{V}^{-1}(\underline{z}-\underline{\eta}_2)$
3	$(\underline{x}-\underline{\mu}_3)'\underline{\Sigma}^{-1}(\underline{x}-\underline{\mu}_3)$	$(\underline{z}-\underline{\eta}_3)'\underline{V}^{-1}(\underline{z}-\underline{\eta}_3)$.

Since the canonical variates are uncorrelated with unit variance, the covariance $\underline{V}$ is in fact the identity matrix $\underline{I}$. The terms $(\underline{x}-\underline{\mu}_i)\underline{\Sigma}^{-1}(\underline{x}-\underline{\mu}_i)$ define equal frequency ellipses on which classification may be based, and the corresponding ellipses in canonical space are given by $(\underline{z}-\underline{\eta}_i)'\underline{V}^{-1}(\underline{z}-\underline{\eta}_i)$. Equal covariances assure identical rankings of observations in both observation and canonical space.

Unfortunately none of these characteristics obtains when covariances are unequal. In particular, the identity of rankings in canonical and observation space no longer holds. This lack of invariance may be indicated by

Ranking	Observation Space	Canonical Space
1	$(\underline{x}-\underline{\mu}_1)'\underline{\Sigma}_1^{-1}(\underline{x}-\underline{\mu}_i)$	$(\underline{z}-\underline{\eta}_i)'\underline{V}_i^{-1}(\underline{z}-\underline{\eta}_i)$
2	$(\underline{x}-\underline{\mu}_2)'\underline{\Sigma}_2^{-1}(\underline{x}-\underline{\mu}_2)$	$(\underline{z}-\underline{\eta}_j)'\underline{V}_j^{-1}(\underline{z}-\underline{\eta}_j)$
3	$(\underline{x}-\underline{\mu}_3)'\underline{\Sigma}_3^{-1}(\underline{x}-\underline{\mu}_3)$	$(\underline{z}-\underline{\eta}_k)'\underline{V}_k^{-1}(\underline{z}-\underline{\eta}_k)$.

Unequal covariances in observation space are indicated by $\underline{\Sigma}_i$, i=1,2,3. Covariances of the canonical variates are also unequal, since the canonical transform generates variates which are globally uncorrelated but conditionally correlated. $\underline{V}_i$ expresses these conditional correlations. Note that the rankings of Mahalanobis distances in canonical space are <u>not</u> given a priori by their rankings in observation space. Since the canonical transform

$$\underline{z} = \underline{U}\underline{x}$$

maintains rankings when covariances are equal, this shift in rankings is a direct result of covariance heterogeneity. This is shown by noting that conditional covariances for transformed and untransformed variates are related by

$$\underline{V}_i = \underline{U}\ \underline{\Sigma}_i\ \underline{U}',$$

where $\underline{U}$ is the (g-1) x p transform matrix generated in equation (1). Then Mahalanobis distances are given by

$$D_i(\underline{z}) = (\underline{z} - \underline{\eta}_i)'\ \underline{V}_i^{-1}\ (\underline{z} - \underline{\eta}_i)$$

$$= [\underline{U}\ (\underline{x} - \underline{\mu}_i)]'\ (\underline{U}\ \underline{\Sigma}_i\ \underline{U})^{-1}\ [\underline{U}\ (\underline{x} - \underline{\mu}_i)]. \quad (6)$$

The influence of $\underline{U}$ on $D_i(\underline{z}) - D_j(\underline{z})$ clearly depends on the difference between $\underline{U}\ \underline{\Sigma}_i\ \underline{U}'$ and $\underline{U}\ \underline{\Sigma}_j\ \underline{U}'$, which in turn depends on the structures of $\underline{\Sigma}_i$ and $\underline{\Sigma}_j$. Therefore the relative magnitudes of Mahalanobis distances in canonical space are <u>a priori</u> indeterminate.

A measure of the distortion induced by covariance heterogeneity may be defined by means of the function

$$h_{ij}(\underline{x}) = 1 \text{ if } D_i(\underline{x})=D_j(\underline{x}) \text{ or } D_i(\underline{z})=D_j(\underline{z})$$

$$= |(D_i(\underline{x})-D_j(\underline{x}))/(D_i(\underline{z})-D_j(\underline{z}))|$$

otherwise,

where $D_i(\underline{z})$ is given by equation (6) and

$$D_i(\underline{x}) = (\underline{x} - \mu_i)' \underline{\Sigma}^{-1} (\underline{x} - \mu_i).$$

This function assumes only positive values, and when covariances are equal its value is unity for all $\underline{x} \in H$. Also, as covariances become more heterogeneous the dispersion of its values increases correspondingly. Now let $h^*(\underline{x})$ be the maximum value of $1/2(h_{ij}(\underline{x}) + 1/h_{ij}(\underline{x}))$ over all i and j:

$$h^*(\underline{x}) =$$

$$\max[1/2(h_{ij}(\underline{x})+1/h_{ij}(\underline{x})):i=1,\ldots,g,$$

$$j=1\ldots g,\ i\neq j].$$

Then the canonical analysis is defined to be distortion-free (within an ϵ-tolerance) if

$$1 - \epsilon \leq h^*(\underline{x}) \leq 1 + \epsilon$$

over some appropriate proportion of the sampling universe H. Note that this condition will be met as long as relative Mahalanobis distances are approximately maintained in canonical space. Since relative distances are exactly maintained when covariances are equal, $h^*(\underline{x}) = 1$ for all $\underline{x} \in H$ and the transform is completely free from distortion.

It should be emphasized that a canonical analysis is distortion-free only for certain combinations of covariance matrices $\underline{\Sigma}_i$, i=1...,g, and that the property cannot be automatically assumed. When the canonical procedure is not distortion-free, statistical relationships between distances in observation space and their canonical representations are complex and non-intuitive. Under such conditions one cannot base inferences about observation data on statistical characterizations in canonical space. Before the canonical analysis can be interpreted with equal frequency ellipses, distortion induced by covariance heterogeneity must be assessed.

A practical consequence for ecologists is that, contrary to the case with equal covariances, one cannot safely use equal frequency ellipses in canonical space for interpreting group differences. The inferences drawn from the canonical analysis are not translatable back to the observation space, because distance measures (and the corresponding probability measures) are distortted by the canonical transform. One conclusion seems inescapable: the canonical analysis, which possesses so many positive geometric and statistical properties for homogeneous covariances, is fraught with problems and ambiguities otherwise. Any recommendation to use this methodology when covariances are demonstrably unequal should be cautiously offered, and cautiously accepted. While heterogeneity of covariance may in fact result in no major misinterpretations, this is by no means a certainty. The point of this discussion is that the canonical analysis requires specific well-defined assumptions, violations of which may have unforeseen and potentially serious distorting effects. The canonical analysis then becomes an ad hoc procedure for generating linear data transforms which may or may not focus the researcher's attention on meaningful relationships. The hope that it will suggests that there may be value in the methodology, even when its assumptions fail. The same claim of course can be made of any data analysis procedure, or indeed of any activity whatever.

CONCLUSIONS

This discussion has focused on a parametric development of canonical analysis and its relationship to the usual data-based approach suggested by Fisher (1936). Notwithstanding the problems associated with covariance heterogeneity, it is fortuitous that maximum likelihood estimation for normal populations yields the familiar computing forms. This insures consistent, asymptotically efficient estimators. Nevertheless, there remain a number of problems concerning the effects of sampling. For example, nothing has been said about small sample variability, rates of convergence, and possible effects of stratified sampling designs.

Results from simulation studies, further theoretical investigations and many different data analyses provide a fairly bleak picture for the use of canonical analysis as an inferential procedure in ecology. The statistical assumptions which insure that the canonical analysis corresponds to posterior classification are almost never met by ecological data. Frequency distributions are almost always non-normal, usually highly skewed, often bimodal, in a great many cases discrete, and covariances are almost universally heterogeneous. The separate and combined effects of these violations on the canonical analysis are almost totally unknown.

Even when the assumptions are met, small sample stability problems arise. Preliminary simulation results by the author indicate considerable instability of the canonical coefficients, due solely to sampling variability. This instability increases rapidly with increases in numbers of variables and groups, and with decreases in sample size and distances between group means. What this means in practice is that any pattern exhibited by the canonical coefficients may be accidental and therefore of no ecological consequence. The effects of sampling variability on the canonical coefficients remain largely unexplored, though the work of Anderson (1963), Rao (1965, 1966) and others provides a theoretical starting point.

Other problems arise when the canonical analysis is based on data gathered by a stratified sampling design. The canonical analysis is sensitive, to a largely unknown degree, to relative as well as absolute sample sizes. This means that decisions concerning relative sampling intensities of groups may have a major impact on the structure of the canonical functions irrespective of the underlying probability structure of the mixture. Such an inherent indeterminancy can undermine any interpretation of the analysis.

There are, in short, a number of more or less serious problms in the application of canonical analysis. That they remain unresolveed at the present time is not an argument to abandon the technique. It does suggest, however, the need for careful planning of studies utilizing canonical analysis, and a healthy scientific skepticism in both the interpretation and the reporting of results.

LITERATURE CITED

Anderson, S.H., and H.H. Shugart, Jr. 1974. Habitat selection of breeding birds in an east Tennessee deciduous forest. Ecology 55:828-837.

Anderson, T.W. 1963. Asymptotic theory for principal component analysis. Annals of Mathematical Statistics 34:122-148.

Bertin, R.I. 1977. Breeding habitats of the wood thrush and veery. Condor 79:303-311.

Buechner, H.K., and H.D. Roth. 1974. The lek system in Uganda kob antelope. American Zoologist 14:145-162.

Buzas, M.A. 1967. An application of canonical analysis as a method for comparing faunal areas. Journal of Animal Ecology 36:563-577.

Cody, M.L. 1978. Habitat selection and interspecific territoriality among the sylviid warblers of England and Sweden. Ecological Monographs 48:351-396.

Conley, W. 1976. Competition between Microtus: a behavioral hypothesis. Ecology 57:224-237.

Conner, R.N., and C.S. Adkisson. 1976. Discriminant function analysis: a possible aid in determining the impact of forest management on woodpecker nesting habitat. Forest Science 22:122-127.

Conroy, M.J., L.W. Gysel, and E.K. Dudderer. 1979. Habitat components of clear-cut areas for snowshoe hares in Michigan. Journal of Wildlife Management 43:680-690.

Dueser, R.D., and H.H. Shugart, Jr. 1978. Microhabitats in a forest-floor small-mammal fauna. Ecology 59:89-98.

Dueser, R.D., and H.H. Shugart, Jr. 1979. Niche pattern in forest-floor small-mammal fauna. Ecology 60:108-118.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Annuals Eugenics 7:179-188.

Goldstein, R.A., and D.F. Grigal. 1972. Definition of vegetation structure by canonical analysis. Journal of Ecology 60:277-284.

Graybill, F.A. 1976. Theory and application of the linear model. 704 p. Duxbury Press, N. Scituate, Mass.

Green, R.H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of central Canada. Ecology 52:543-556.

Green, R.H. 1974. Multivariate niche analysis with temporally varying environmental factors. Ecology 55:73-83.

Grigal, D.F., and R.A. Goldstein. 1971. An integrated ordination-classification analysis of an intensively sampled oak-hickory forest. Journal of Ecology 59:481-492.

Harner, E.J., and R.C. Whitmore. 1977. Multivariate measures of niche overlap using discriminant analysis. Theoretical Population Biology 12:21-36.

Horton, I.F., J.S. Russell, and A.W. Moore. 1968. Multivariate-covariance and canonical analysis: a method for selecting the most effective discriminators in a multivariate situation. Biometrics 24:845-858.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Kowal, R.R., M.J. Lechowicz, and M.G. Adams. 1972. The use of canonical analysis to compare response curves in physiological ecology. Flora 165:29-46.

MacArthur, R.H., and R. Levins. 1967. The limiting similarity, convergence, and divergence of coexisting species. American Naturalist 101:377-385.

Montanucci, R.R. 1978. Discriminant analysis of hybridization between leopard lizards, Gambelia (Reptilia, Lacertilia, Iguanidae). Journal of Herptology 12:299-307.

Norris, J.M., and J.P. Barkham. 1970. A comparison of some Cotswold beechwoods using multiple-discriminant analysis. Journal of Ecology 58:603-619.

Rao, C.R. 1965. Linear statistical inference and its applications. 522 p. John Wiley and Sons, New York, N.Y.

Rao, C.R. 1966. Inference on discriminant function coefficients. p. 587-602. In Bose, R.C., I.M. Chakravarti, P.C. Mahalanobis, C.R. Rao, and K.J.C. Smith, editors. Essays in probability and statistics. University of North Carolina Press, Chapel Hill, N.C.

Rice, J.C. 1978a. Behavioral interactions of interspecifically territorial vireos. I. Song discrimination and natural interactions. Animal Behaviour 26:527-549.

Rice, J.C. 1978b. Behavioral interactions of interspecifically territorial vireos. II. Seasonal variation in response intensity. Animal Behaviour 26:550-561.

Rice, J.C. 1978c. Ecological relationships of two interspecifically territorial vireos. Ecology 59:526-538.

Ricklefs. R.E. 1977. A discriminant function analysis of assemblages of fruit-eating birds in Central America. Condor 79:228-231.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58:810-819.

Toussaint, G.T. 1974. Bibliography on estimation of misclassification. IEEE Transactions Inferential Theory. IT-20:472-479.

Weiner, J.M., and O.J. Dunn. 1966. Elimination of variates in linear discrimination problems. Biometrics 22:268-275.

DISCUSSION

LESLIE MARCUS: Your approach requires that each sample of the habitat correspond to a unique species utilizing it. In some restricted cases this is appropriate, but in a great many others it is not. More generally there is some probability of use by each of the species, so that a technique like canonical correlation is more appropriate.

KEN WILLIAMS: I agree. The conceptual framework of DA requires the existence of distinct statistical populations and an unambiguous association of sampling units to them. There is a very large number of problems which fit this framework, as indicated in my paper. However, multiple use of sampling units by different species does not. Problems of species-habitat or community-habitat relationships for such cases could better be handled by something like canonical correlation. Another possibility might be to combine correspondence analysis and multivariate multiple regression.

BARRY NOON: There are instances in the literature where principal components analysis (PCA) and discriminant analysis (DA) have been used with the same set of data. How does one interpret differences in ordination loadings?

KEN WILLIAMS: These two procedures focus on different structural features of the data, and address different questions. It is important to recognize this, in order to avoid misinterpretation of the different results. Some confusion about the relationship of PCA and DA probably arises from their mathematical similarities. Both, for example, involve an eigenstructure analysis, and both can be used to assess differences among groups of observations in a space of reduced dimensionality. In the case of PCA the space is defined by the dominant eigenvectors of a covariance (or correlation) matrix, whereas for DA the space is generated from eigenvectors of a matrix involving group means themselves. Nonetheless, both techniques represent a coordinate rotation of multivariate measurements and a reduction of their dimensionality.

The basic difference between them is in the way the group structure is accounted for. DA requires well defined groups, and utilizes this feature by means of the equation

$$[\underline{B} - \lambda \underline{W}]\underline{u} = 0.$$

PCA, on the other hand, uses the data matrix without reference to structure:

$$[\underline{T} - \lambda \underline{I}]\underline{v} = 0$$

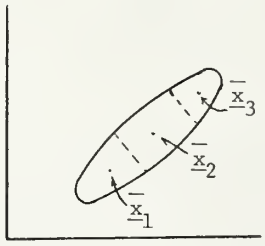
or

$$[\underline{R} - \lambda \underline{I}]\underline{v} = 0,$$

where $\underline{R}$ is the correlation matrix based on $\underline{T}$, and

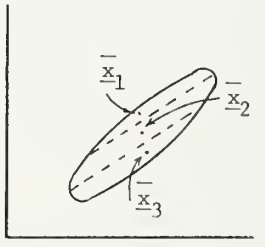
$$\underline{T} = \underline{B} + \underline{W}.$$

Clearly the relationship of axes for PCA and DA depends on differences in structure for $\underline{T}$, $\underline{B}$ and $\underline{W}$. To see how these differences may affect interpretation of the data, consider a partition of data into groups such that the group means lie along the dominant axes of dispersion. In two dimensions this situation appears as:



Then provided $\lambda_i / \sum_i \lambda_i$ is close to unity (i.e., the first eigenvector dominates the eigenstructure of T) the principal component $\underline{v}_i = \underline{P}_1' \underline{x}$ and canonical variate $\underline{z}_1 = \underline{u}_1' \underline{x}$ will be approximately the same linear function. Under these conditions the clustering of groups will be displayed in component as well as discriminat space.

Now consider a partition which looks like:



The dominant principal component will not display clustering of groups in this situation, since groups overlap extensively along the dominant axis. DA, however, displays almost total separation of groups along the dominant canonical axis, reflecting the data partition. This difference between principal components and the canonical functions demonstrates that group structure, though highlighted by appropriate linear transforms (the canonical variates), may be completely obscured by others (the principal components).

Thus there is no guarantee that PCA will display important data structures; nor can one conclude a lack of structure based on projections of data in component space. This is not unexpected, since PCA is designed specifically for variance maximization rather than maximum group separation. In fact, every linear transform of observation data gives us a different look at the multivariate system. What one sees depends in large measure on the manner in which he looks, i.e., the transform he uses. PCA provides insight into the components of overall variation for a system, but not necessarily its group structure. DA gives a look at group separation, but not at overall variation. That these two procedures often ordinate in quite different ways therefore should come as no surprise.

THEORY AND METHODS OF FACTOR ANALYSIS AND PRINCIPAL COMPONENTS[1]

Helen Bhattacharyya[2]

Abstract.--A brief discussion is given on the historical development of factor analysis, the kind of data for which factor analysis is applicable, and the kind of result one may obtain. The distinction between principal components analysis and factor analysis is clarified, and the relationship between the unique factors and the communalities in the reduced correlation matrix described. A derivation of the principal components is given, followed by a description of the principal factor solution and interated principal factor solution. Without mathematical details, the concept of maximum likelihood solution is introduced. Once a direct solution is obtained possible rotations, orthogonal and oblique, are discussed as attempts at deriving more conceptually meaningful common factors. The nonuniqueness of factorization of the correlation matrix is shown. The concept of simple structure is introduced and a graphical representation is given illustrating the meaning of rotation of factors. Factor scores and scoring coefficient matrix are described.

Key words: Characteristic equation; characteristic root; communalities; correlation matrix; factor analysis; least squares; maximum likelihood; principal components.

INTRODUCTION

In understanding factor analysis, it helps to know a few things that the procedure does not do. Unlike discriminant analysis, factor analysis does not attempt to distinguish two or more distinct populations. There is no classification problem. Unlike multiple regression, factor analysis does not estimate any relationship between a set of independent variables and one or more dependent variables. The kind of data for which factor analysis is applicable usually consists simply of one sample of multivariate observations. With a sample size N and n variables, the raw data for factor analysis would be of the following form:

Obs.	Z_1	Z_2	. . . . .	Z_n
1				
2				
.				
.				
.				
N				

where Z_1, Z_2, ..., Z_n denote the n variables measured for each observation. An example of the variables Z_i, i = 1, ..., n, follows:

Eight Physical Measurements (Harman 1968)

Z_1 = Height
Z_2 = Arm span
Z_3 = Length of forearm

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Mathematical Statistician, USDA Forest Service, Southeastern Forest Experiment Station, Research Triangle Park, NC 27709.

Z_4 = Length of lower leg
Z_5 = Weight
Z_6 = Bitrochanteric diameter
Z_7 = Chest girth
Z_8 = Chest width.

The primary purpose of factor analysis is to bring about a reduction in dimensionality, to explore relationships among large numbers of observed variables in an effort to find "factors" that reduce the complexity of the situation. In the example, the goal is to find a relatively few underlying common factors such that the eight measurements may be described as a linear function of these common factors.

HISTORICAL BACKGROUND

Spearman (1904) first introduced the concept of factor analysis in an article titled "General intelligence, objectively determined and measured." Early researchers in the field were primarily concerned with finding one common underlying factor, intelligence for example, that would help explain an individual's performance on a variety of different tests. Spearman called this the two-factor theory, one common factor and another specific to a particular test. By the early 1930's it became evident, through the writings of Thurstone (1931) and others, that the two-factor theory was not sufficient to describe a battery of psychology tests. In the 1930's, there developed the concept of multifactor theory; that is, there can be more than just one common factor. The subject of factor analysis, then, is to find the common factors and the relationship between the observed variables and the common factors.

FACTOR ANALYSIS MODEL

Let the observed variables be denoted by Z_1, Z_2, ..., Z_n and the common factors by F_1, F_2, ..., F_m where $m < n$. With the inclusion of the unique factors, U_1, U_2, ..., U_n, the basic linear relationship between variables and factors may be written:

$$Z_1 = a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + d_1U_1$$

$$Z_2 = a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + d_2U_2$$

.

.

.

$$Z_n = a_{n1}F_1 + a_{n2}F_2 + \ldots + a_{nm}F_m + d_nU_n ,$$

where a_{ij} and d_i, $i=1,\ldots,n$, $j=1,\ldots,m$, are constants. The matrix $((a_{ij}))$ is called the factor pattern matrix, and an element a_{ij} is called the factor loading of variable Z_i on factor F_j. The unique factors U_i are assumed to be independent of the common factors and independent of each other. For simplicity of exposition, the common factors may be assumed orthogonal to each other, although this is not strictly necessary as will be seen in the discussion of oblique factor rotation. A further assumption is that Z_i, F_j, U_i are all standardized to have mean 0 and unit variance.

Referring to the above example, Z_i would denote the observed variables height, arm span, ..., chest width, so that n=8. The F_j, $j=1,\ldots,m$, are the common factors. Prior to performing the factor analysis the common factors and the value of m are assumed unknown. Letting R denote the observed correlation matrix corresponding to the N n-variate observations, the problem is to find a numerical solution for the elements of the pattern matrix $((a_{ij}))$ which would, in some sense, best reproduce R.

PRINCIPAL COMPONENT ANALYSIS

The use of principal components as a data reduction technique was introduced by Pearson (1901) and further developed by Hotelling (1933). The principal component solution will be described first as it avoids some of the inherent difficulties in factor analysis.

Principal Component Model

Let the principal components be denoted as P_1, P_2, ..., P_n. The linear relationship between Z_i, the observed variables, and P_j may be written:

$$Z_1 = a_{11}P_1 + a_{12}P_2 + \ldots + a_{1n}P_n$$

$$Z_2 = a_{21}P_1 + a_{22}P_2 + \ldots + a_{2n}P_n$$

.

.

.

$$Z_n = a_{n1}P_1 + a_{n2}P_2 + \ldots + a_{nm}P_n .$$

The principal components P_j, $j=1,\ldots,m$, where $m=n$, are assumed to be orthogonal to each other and to have variances $\mathrm{Var}(P_1) \geq \mathrm{Var}(P_2) > \ldots \geq \mathrm{Var}(P_n)$.

Note that this model differs from the factor analysis model in that $m = n$ and that there are no unique factors. As in factor analysis, the objective is to find the matrix of coefficients $((a_{ij}))$.

Principal Component Solution

Consider the derivation for P_1, the first principal component. Let $P_1 = \underline{a}'\underline{Z} = a_1Z_1 + a_2Z_2 + \ldots + a_nZ_n$, then

$$\mathrm{Var}(P_1) = \underline{a}'R\underline{a}, \qquad (1)$$

where R is the covariance matrix of the Z_i's, or equivalently the correlation matrix of the Z_i's if all Z_i's are standardized to mean 0 and unit variance. Although standardizing the Z_i's is not necessary in principal component analysis, this is usually recommended especially when the variables are measured in different scales.

The next step is to maximize $\mathrm{Var}(P_1)$, subject to a normalizing restraint $\underline{a}'\underline{a} = 1$. Using Lagrange multiplier λ and setting to 0 the partial derivatives with respect to a_i,

$$\frac{\partial}{\partial \underline{a}}\left[\underline{a}'R\underline{a} + \lambda(1 - \underline{a}'\underline{a})\right] = 0$$

$$2(R - \lambda I)\underline{a} = \underline{0}. \qquad (2)$$

For a nontrivial solution to (2) λ must satisfy the characteristic equation

$$|R - \lambda I| = 0 \qquad (3)$$

The left hand side of (3) is a polynomial in λ of degree n, hence there are n solutions $\lambda_1, \lambda_2, \ldots, \lambda_n$. These are called the characteristic roots (or eigenvalues), and corresponding to each λ_i there is a characteristic vector (or eigenvector) $\underline{a}_i$ such that

$$(R - \lambda_i I)\underline{a}_i = \underline{0}.$$

Premultiplying (2) by $\underline{a}'$ gives $\underline{a}'R\underline{a} - \underline{a}'\lambda I\underline{a} = 0$, or $\underline{a}'R\underline{a} = \lambda$. But, $\underline{a}'R\underline{a} = \mathrm{Var}(P_1)$ from (1). Therefore, $\mathrm{Var}(P_1) = \lambda_1$, where λ_1 denotes the largest characteristic root; and $P_1 = \underline{a}_1'\underline{Z}$, where $\underline{a}_1$ is the characteristic vector corresponding to λ_1.

Similarly, the second principal component $P_2 = \underline{b}'\underline{Z}$ may be derived using restriction $\underline{b}'\underline{b} = 1$ and the further restriction $\underline{a}_1'\underline{b} = 0$. However, this is not necessary. If the characteristic roots $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are ordered, then the ith principal component is $P_i = \underline{a}_i'\underline{Z}_i$ and $\mathrm{Var}(P_i) = \lambda_i$.

One may ask what has been accomplished. The n variables have resulted in n principal components. No apparent reduction in dimensionality has been achieved if all principal components are kept. However, in defining the principal components, it is hoped that the sum of the m characteristic roots, $m < n$, would nearly equal n. The sum of the variances of Z_i is $\sum_{i=1}^{n} \mathrm{Var}(Z_i) = n \times 1 = n$ and the sum of the variances of the principal components is $\sum_{i=1}^{n} \mathrm{Var}(P_i) = \sum_{i=1}^{n} \lambda_i = n$. If $\sum_{i=1}^{m} \lambda_i$ is very near n, then the components corresponding to the smaller λ_i, $i = m + 1, \ldots, n$, contribute very little to the variation in the Z_i and, hence, may be neglected.

FACTOR ANALYSIS

Recall the factor analysis model given earlier,

$$Z_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{im}F_m + d_iU_i, \quad i=1,\ldots,n,$$

where F_i are orthogonal to each other and independent of U_i,

$$\mathrm{Var}(Z_i) = 1 = \mathrm{Var}\left(\sum_{j=1}^{m} a_{ij}F_j\right) + \mathrm{Var}(d_iU_i).$$

Let $h_i^2 = 1 - \mathrm{Var}(d_iU_i)$. The quantity of h_i^2, $i=1,\ldots,n$, is called the communality of Z_i because it represents the variance of Z_i due to the common factors F_j, $j=1,\ldots,m$.

In solving for the factor pattern matrix, the object is to find that set of loadings a_{ij} which best reproduces the correlation matrix. Since the variance of Z_i due to the common factors F_j, $j=1,\ldots,m$, is h_i^2 rather than unity, it is reasonable then to substitute h_i^2 for unity in the diagonal elements of the correlation matrix. The result is called the reduced correlation matrix and computationally this constitutes the basic difference between the principal component solution and the common factor solution. Note that in principal component analysis the variance of Z_i due to the components is necessarily unity since there is no provision for a unique factor.

Two commonly used methods for estimating the communalities are 1) take h_i^2 to be the square of the multiple correlation coefficient between Z_i and the other n-1 Z's and 2) take h_i^2 to be the largest (absolute value) correlation in each row.

Principal Factor Solution

Once the numerical value of each element

including the communalities in the reduced correlation matrix is obtained, the solution for the factor pattern matrix may proceed as with the principal components solution. The characteristic equation is formed and solved for the characteristic roots and characteristic vectors. The reduced correlation matrix for the example is

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Z_1	h_1^2		(symmetric)					
Z_2	0.846	h_2^2						
Z_3	0.805	0.881	h_3^2					
Z_4	0.859	0.826	0.801	h_4^2				
Z5	0.473	0.376	0.380	0.436	h_5^2			
Z_6	0.398	0.326	0.319	0.329	0.762	h_6^2		
Z_7	0.301	0.277	0.237	0.327	0.730	0.583	h_7^2	
Z8	0.382	0.415	0.345	0.365	0.629	0.577	0.539	h_8^2

At this point, the question that usually arises is how many factors should be retained. Although under certain normality conditions the statistical significance for the number of factors can be tested, the usual rule of thumb is to keep only those characteristic roots of value greater than unity. This concept is intuitively reasonable since $\text{Var}(Z_i) = 1$, and any common factor which accounts for less than unit variance would not be very helpful in achieving parsimony.

A number of other factorization methods are available. Two other methods will be described briefly, the iterated principal factor solution and the maximum likelihood solution.

Iterated Principal Factor Solution

This method may be considered an improved principal factor solution. Beginning with initial estimates of communalities, factor loadings are obtained by the principal factor method. Using the factors extracted, new estimates of communalities are computed and new factor loadings obtained. This iterative procedure is continued until the new communalities differ by less than a predetermined small amount from the previous communalities.

Maximum Likelihood Solution

Although the principal factor solution was well accepted by psychologists and social scientists, there were mathematical statisticians who continued trying to find a solution within the rigorous framework of statistical inference. The breakthrough came when Lawley (1940) developed equations for the maximum likelihood estimation of factor loadings. Although the derivation and algorithms are extremely difficult, the basic concepts involved are those of standard maximum likelihood estimation.

The variables Z_i, $i=1,\ldots,n$, are assumed to follow an n-variate normal distribution, hence the sample covariance matrix follows a Wishart distribution. (Recall in the principal factor solution no assumptions were made on the underlying distribution of the variables Z_i.) The goal is to find the expression for the loadings which would maximize the Wishart density. No assumptions are made about the communalities, but it is necessary to assume the number of common factors. This number, however, may be tested subsequently for goodness of fit using a likelihood ratio test. The maximum likelihood solution for factor loadings has the usual desirable properties of maximum likelihood estimators, such as consistency, asymptotic efficiency, and asymptotic normality. One disadvantage of the method at the present time is that it does not always converge properly depending on the particular data set and the computer package used.

ROTATION OF FACTORS

The ability to describe a set of n variables in terms of m factors implies that the observed data points essentially lie in an m-dimensional space imbedded in the original n-dimensional space. However, there is not a unique set of axes, or factors, for describing this m-dimensional space. It can be shown algebraically that the factorization of the correlation matrix is not unique.

The factor analysis model presented earlier may be written as $\underline{Z} = A\underline{F}$ if we omit the unique factors and use A to denote the n x m pattern matrix. Assuming the common factors to be orthogonal and standardized, the variances and covariances of the standardized variables $Z_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{im}F_m$, $i=1,\ldots,n$, may be denoted as

$$\text{Var}(Z_i) = r_{ij} = \sum_{j=1}^{m} a_{ij}^2 \text{ and}$$

$$\text{Cov}(Z_i Z_k) = r_{ik} = \sum_{j=1}^{m} a_{ij}a_{kj}$$

so that the correlation matrix of Z_i may be written $R = AA'$. If T is an orthonormal matrix and the transformation $B = AT$ is made, then $BB' = (AT)(AT)' = ATT'A' = AIA' = R$. Factorization of R is not unique since it may be factored into AA' as well as BB'. The question, then, is which factor pattern matrix should be chosen, A or B.

Intuitively, it is obvious that the factor pattern chosen should be the one which leads to conceptually meaningful factors. Thurstone (1935, 1947) gave the following rules which he called

simple structure principles for choosing the factor pattern matrix.

1. Each row of the factor matrix should have at least one zero.
2. If there are m common factors, each column of the factor matrix should have at least m zeros.
3. For every pair of columns of the factor matrix there should be several variables whose entries vanish in one column but not in the other.
4. For every pair of columns of the factor matrix, a large proportion of the variables should have vanishing entries in both columns when there are four or more factors.
5. For every pair of columns of the factor matrix, there should be only a small number of variables with non-vanishing entries in both columns.

The ideas embodied in these rules, simplification of the rows and simplification of the columns, were formalized by Carrol (1952), Kaiser (1958), and others and led to the orthogonal and oblique rotations commonly used today.

Quartimax Rotation

This is an orthogonal transformation which attempts to simplify the rows of the pattern matrix. The quantity maximized is

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^4. \qquad (4)$$

The name quartimax is applied because a_{ij} is raised to the 4th power. This procedure is equivalent to minimizing Q' where

$$Q' = \sum_{i=1}^{n} \sum_{j<\ell=1}^{m} (a_{ij}a_{i\ell})^2. \qquad (5)$$

It is evident from (5) that minimizing Q' simplifies the rows of the pattern matrix. It will be shown now that maximizing Q is equivalent to minimizing Q'.

Since communalities, and hence the squares of communalities, remain constant under orthogonal transformations, we have

$$(\sum_{j=1}^{m} a_{ij}^2)^2 = \sum_{j=1}^{m} a_{ij}^4 + 2 \sum_{j<\ell=1}^{m} (a_{ij}a_{i\ell})^2 = \text{constant}.$$

Summing over the n variables gives

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^4 + 2 \sum_{i=1}^{n} \sum_{j<\ell=1}^{m} (a_{ij}a_{i\ell})^2 = \text{constant}.$$

Since the sum of the two terms on the left hand side is a constant, maximizing the first term is equivalent to minimizing the second term.

Varimax Rotation

This orthogonal transformation attempts to simplify the columns of the pattern matrix. If a simple factor is defined as one with only 0's and 1's in the column, then the varimax rotation attempts to bring about this simplicity by maximizing V_j.

$$V_j = 1/n \sum_{i=1}^{n} (a_{ij}^2)^2 - 1/n^2 (\sum_{i=1}^{n} a_{ij}^2)^2,$$

$j=1, 2, \ldots, m$, the variance of the squared loadings in each column. Maximizing $\sum_{j=1}^{m} V_j$ results in a rotation overly influenced by the size of the communalities of the variables. In practice, this is taken into account and the quantity actually maximized in the varimax rotations is

$$n^2V = n \sum_{j=1}^{m} \sum_{i=1}^{n} (a_{ij}/h_i)^4 - \sum_{j=1}^{m} (\sum_{i=1}^{n} (a_{ij}/h_i)^2)^2.$$

Oblique Rotation

The criterion of simple structure remains as with orthogonal rotations, but the method is no longer limited to orthogonal transformations of the factors. The use of oblique factors is becoming increasingly popular because it affords more flexibility in defining the underlying factors and because computational difficulties are no longer an obstacle with the availability of computer packages. Harman (1968) gives an excellent discussion of several oblique rotation methods.

Graphical Representation

Thurstone's (1935, 1947) simple structure principles may be interpreted graphically as requiring the data points to lie, to the extent possible, on or near the reference axes defined by the factors. Factor rotation then is an effort to rotate the axes to achieve this goal. Figure 1 gives an illustration of orthogonal and oblique rotations for two factors. Data points are denoted by X's. The original factor solution is represented by F_1, F_2, the rotated orthogonal solution by G_1, G_2, and the rotated oblique solution by G_1 and G_2'.

Numerical Example

The factor patterns under different rotations for the example is given in table 1. Two factors were retained in the initial solution. Under all

rotations, variables Z_1 to Z_4 loaded heavily on F_1 and variables Z_5 to Z_8 loaded heavily on F_2. An examination of these variables leads us to conclude that Factor 1 is a measure of lankiness and Factor 2 is a measure of stockiness.

FACTOR SCORES

Often the interest in factor analysis does not end in simply being able to obtain the factor pattern and verbalize the factors extracted. Sometimes one may wish to go one step further and find the "observed" values, or scores, of the factors; that is, to describe the factors in terms of the observed variables. This may be useful, for example, in carrying out a regression analysis using m factors instead of the original n variables as the independent variables.

In principal component analysis, the computation of the scoring coefficient matrix is straightforward. In the model $\underline{Z} = A\underline{P}$, the matrix A is square and usually full rank, and the solution for $\underline{P}$ is simply $\underline{P} = A^{-1}\underline{Z}$.

For the factor analysis model,

$$Z = A\underline{F} + D\underline{U}, \qquad (6)$$

where the matrix A is n x m and D denotes the diagonal matrix containing the coefficient d_i of the unique factors, a least squares regression approach is generally used. If $D\underline{U}$ is the error term, then (6) may be considered a linear regression of $\underline{Z}$ on A where $\underline{F}$ are the coefficients. From standard regression theory the least squares solution for $\underline{F}$ is $\underline{F} = (A'A)^{-1}A'\underline{Z}$.

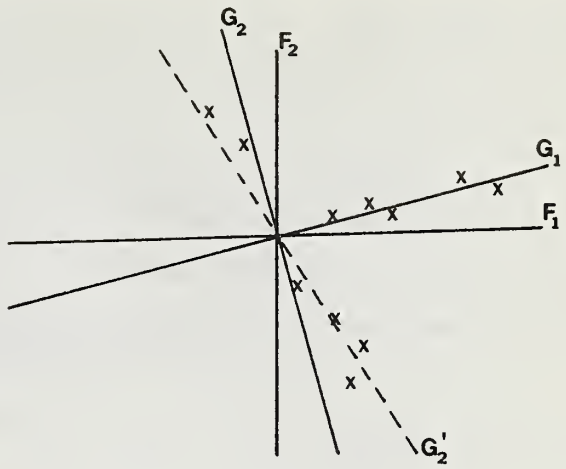


Figure 1. Graphical representation of factor rotation illustrating original factor solution (F_1,F_2); rotated orthogonal solution (G_1,G_2); and rotated oblique solution (G_1,G'_2).

COMPUTER PACKAGES

Computer programs for factor analysis may be found in SAS (Statistical Analysis Systems 1979), SPSS [Statistical Package for the Social Sciences (Kim 1975)], and other statistical packages. Typically, the programs are easy to use and only require a few lines of code. In SAS, for example, the following code would produce a factor analysis using the iterated principal factor method and the varimax rotation on variables Z_1, Z_2, ..., Z_8.

```
PROC FACTOR METHOD=PRINIT ROTATE=VARIMAX SCORE;
     VAR Z1 - Z8;
```

Table 1. Factor patterns for example on eight physical measurements.

	Factor pattern							
	Initial solution		Quartimax rotation		Varimax rotation		Oblique rotation	
	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
Z_1	0.86	-0.33	0.90	0.20	0.88	0.27	0.78	0.05
Z_2	0.85	-0.41	0.93	0.13	0.92	0.21	0.84	-0.03
Z_3	0.81	-0.41	0.90	o.10	0.89	0.18	0.81	-0.05
Z_4	0.83	-0.34	0.88	0.17	0.86	0.25	0.77	0.02
Z_5	0.75	0.56	0.32	0.88	0.24	0.90	-0.00	0.83
Z_6	0.64	0.51	0.25	0.77	0.18	0.79	-0.01	0.71
Z_7	0.56	0.49	0.20	0.72	0.13	0.73	-0.06	0.70
Z_8	0.62	0.37	0.31	0.66	0.25	0.68	0.09	0.57

Output would include the correlation matrix, characteristic roots, the factor pattern matrix, the rotated factor pattern matrix, and the scoring coefficient matrix. To create a new data set containing the factor scores for the original observations, the above code may be modified.

```
PROC FACTOR METHOD=PRINIT ROTATE=VARIMAX
     SCORE OUT=COEFF; VAR Z1 - Z8;

PROC SCORE DATA=OLD SCORE=COEFF OUT=NEW;
```

In this example OLD is the name of the original data set containing the N observations on variables Z_1, Z_2, ..., Z_8, and NEW is the name of the newly created data set containing the N factor scores on the factors F_1, F_2, ..., F_m.

LITERATURE CITED

Carrol. J.B. 1953. An analytical solution for approximating simple structure in factor analysis. Psychometrika 18:23-28.

Harman, H.H. 1968. Modern factor analysis. Second edition. 469 p. University of Chicago Press, Chicago.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Education Psychology 24:417-41, 498-520.

Kaiser, H.F. 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23:187-200.

Kim, J. 1975. Factor analysis, p. 469-514. In Nie, N.H., et al., editors, SPSS: statistical package for social sciences. 675 p. McGraw-Hill Co., New York, N.Y.

Lawley, D.N. 1940. The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh 60:64-82.

Pearson, K. 1901. On lines and points of closest fit to systems of points in space. Philosophical Magazine 6:559-72.

SAS Institute, Inc. 1979. SAS User's Guide. 494 p. SAS Institute, Inc., Raleigh, N.C.

Spearman, C. 1904. General intelligence, objectively determined and measured. American Journal of Psychology 15:201-93.

Thurstone, L.L. 1931. Multiple factor analysis. Psychology Review 38:406-27.

Thurstone, L.L. 1935. The vectors of mind. 266p. University of Chicago Press, Chicago.

Thurstone, L.L. 1947. Multiple factor analysis. 535 p. University of Chicago Press, Chicago.

DISCUSSION

JAMES DUNN: I question the statement that PCA does not assume anything. Certainly the existence of second-order moments, i.e., a covariance matrix, is assumed. It seems nonsense to apply PCA to this if the variances and covariances depend on the mean values, i.e., as in sampling from multivariate Poisson distribtution. But the only distribution in which the covariance matrix is not functionally related to the mean is multinormal.

HELEN BHATTACHARYYA: Principal components and factor analysis are performed on the sample covariance or sample correlation matrix, which always exists with any reasonable data regardless of the existence of the population moments. Classical factor analysis is a method of determining the dimensionality of the data space and does not attempt at statistical inference or tests of hypotheses, procedures for which assumptions on underlying distributions are necessary. There is no disagreement on the assumption of multivariate normal distribution when the maximum likelihood method of factorization is used.

As to the functional dependence of the mean and the variance, surely the variance of a uniform distribution on (c-a, c+a) is not functionally dependent on the mean c.

JAMES DUNN: What is your impression of the usefulness of oblique factor solution in habitat analysis?

HELEN BHATTACHARYYA: Since our interest is in extracting factors that are conceptually meaningful rather than mathematically expedient, there is no reason to limit ourselves only to orthogonal factors.

KEN MORRISON: Is there a reason why you did not mention equamax rotation?

HELEN BHATTACHARYYA: Equamax rotation is somewhat of a compromise between quartimax rotation and varimax rotation and may be considered very desirable. There was no reason why it wasn't mentioned, other than that a number of other rotation methods were also not mentioned.

MICHAEL KINGSLEY: Principal components are known not to be invariant against linear transforms of data. Is factor analysis free of similar quirks?

HELEN BHATTACHARYYA: It is true that PCA when performed on the variance covariance matrix is not invariant under linear transformations of the variables. Factor analysis is performend on the correlation matrix and does remain invariant as correlations remain invariant under linear transformations.

JAMES SKALEY: When using PCA or factor analysis there can be non-linear relationships in the data set along with unequal covariances. The results have been distorted, and we cannot assume the underlying linear model. Are there guidelines for the interpretation of the results from these analysis?

HELEN BHATTACHARYYA: Factor analysis assumes the observed variables may be expressed as a linear combination of the factors. If this is not true then the methodologies described are not appropriate. If factor analysis is performed, I do not believe there can be a general guideline on interpreting the results. For a discussion on nonlinear reduction of dimensionality see R. Gnanadesikan (1977. Methods for statistical data analysis of multivariate observations. John Wiley & Sons, New York, N.Y.).

B. KEN WILLIAMS: Just a comment: there are both scale-free and non-scale-free techniques for deteriming factor loadings. A good description is found in N.H. Timm (1975. Multivariate analysis with applications in education and psychology. Brooks-Cole, Monterey, Calif.).

JAMES HARNER: This is merely a comment. I see a fundamental difference between principal components analysis (PCA) and factor analysis. PCA is simply a rotation of axes to achieve certain criteria, whereas factor analysis is based on a model. Also the question of measurement scale is important in ecology. A PCA can be done by scaling with other means than just using the correlation matrix.

CANONICAL CORRELATION ANALYSIS AND ITS USE IN WILDLIFE HABITAT STUDIES[1]

Kimberly G. Smith[2]

Abstract.--Canonical correlation is the generalized case of multiple regression wherein one attempts to examine interrelationships between two (or more) sets of variables simultaneously. This is accomplished by maximizing the correlation between a composite of variables from one set with a composite of variables from the second set. The statistical assumptions involved are discussed briefly and a geometric representation of the procedure is presented.

A major problem with canonical correlation analysis has been interpretation of results. Cross-validation aids in detection of sample-specific covariation to which canonical correlation is quite sensitive. The test of redundancy offers a method whereby relationships between data sets themselves can be examined rather than relationships between composites of two data sets by calculating the variance of one set that is accounted for by the second set. Selection of variables is critically important; sample sizes should be as large as possible; and jackknifing estimators and use of robust estimators may be helpful.

Canonical correlation analysis has been used most often in the fields of education and psychology. It has met with limited success in phytosociological studies partly because the requirement of linear relationships between variables is too strict. In ecological research, studies have sought to characterize morphological and behavioral data with environmental data. Most animal community studies that have used canonical correlation have suffered from small sample sizes. Canonical correlation analysis seems to hold some promise for wildlife habitat studies, and its use should increase as researchers become more familiar with it. The need for planning and appropriate data collection methods are extremely important.

Key words: Bartlett's test of significance; canonical correlation analysis; cross-validation; jackknifed estimators; ordination; redundancy; robust estimators.

INTRODUCTION

Ecological researchers often collect data on two sets of variables and attempt to elucidate relationships between the sets. The two data sets may have a cause-effect (predictor-criterion) relationship; but, more commonly, one is interested in relating biotic and abiotic factors,

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington Vt.

[2]Research Ecologist, University of California, Bodega Marine Laboratory, P. O. Box 247, Bodega Bay, CA 94923

e.g., the response of an animal population to a suite of environmental factors along some gradient. Measurements may be from only one area, or encompass data collected simultaneously in several areas. The available statistical options include reporting simple product-moment correlations (r), principal component analysis combining the two sets of variables, or canonical correlation analysis (Barkham and Norris 1970). In most cases, calculating simple correlations is not very informative, and complexity (and usually confusion) increases with the size of the data set. Combining the two data sets into one principal component analysis does not maintain the distinctiveness of the two data sets and is therefore difficult to interpret although it does highlight variables that covary. Canonical correlation analysis, a multivariate statistical technique that attempts to find maximal correlations between two data sets, would appear therefore to hold some promise for wildlife habitat studies. Canonical correlation can also be used as a multiple partial correlation technique (Cooley and Lohnes 1971) and has been employed with limited success as an ordination technique (Gauch and Wentworth 1976).

Canonical correlation analysis was developed by Hotelling (1935, 1936) to examine the relationships between a set of mental test scores and measurements of performance on the same test subjects, although seeds of the technique can be found in the works of Galton, Edgeworth and Pearson at the turn of the century (see Bryant and Atchley [1975] for historical development and comprehensive bibliography). Because calculations involved are extremely laborious to perform by hand, canonical correlation did not become a widely used statistical tool until the advent of computers and publication of a computer program by Cooley and Lohnes (1962). Other problems that have been associated with the use of canonical correlation are the availability of more familiar techniques; the difficulty of interpreting results; and the tendency of results to be situation-specific and not generalizable (Thorndike and Weiss 1973). Canonical correlation has been applied widely in psychological and educational research, but has seen little use in biological research, despite recommendations (e.g., Dunn 1972, Pielou 1977).

Another reason canonical correlation may have been overlooked in biological research is that often biologists are more interested in explaining variance associated with individual measurements (e.g., through the use of principal component analysis) whereas psychological and educational workers have been more interested in correlational relationships (Rao 1955). In some instances, principal component and canonical correlation analyses will produce similar results (e.g., Cassie and Michael 1968, Gauch and Wentworth 1976); however, differences in results from the two procedures may also be of interest (e.g., Barkham and Norris 1970). Cassie (1969) pointed out the great potential for establishing important correlational relationships in biological research, especially when it is tedious and/or expensive to obtain one of the desired data sets. For example, Cassie suggested it would be easier to monitor water temperature and salinity automatically than to collect and analyze plankton samples. Conversely, it would be easier to collect samples of beach infaunal organisms than to determine particle size distributions of the substrate at many sampling sites. The potential value of establishing correlational relationships in such decision-making processes as environmental impact statements or management programs is obvious.

CANONICAL CORRELATION ANALYSIS

Statistical Assumptions

The usual multivariate statistical assumptions apply to canonical correlation analysis: 1) The covariance matrix must be of full rank (Morrison 1976), i.e., there must be more samples than number of variables in both data sets combined. 2) No singularities can exist within the data matrices, i.e., all rows (or columns) must be independent. Singularities arise when a row (or column) of the data matrix is a linear combination of one or more of the other rows (or columns) (see Gauch and Wentworth 1976). Canonical correlation analysis has no solution when a matrix singularity is encountered. The possibility of singularities in large sets of distributional data may limit the usefulness of canonical correlation as an ordinational technique (e.g., Barkham and Norris 1970).

Two other statistical conditions are usually associated with canonical correlation: a multivariate normal (or multinormal) distribution and linear relationships between variables (see Cassie 1969). The constraint of linear relationships between variables (inherent in any linear model, e.g., principal component analysis) limits the usefulness of canonical correlation as an ordination technique for communities, such as in phytosociological studies (Gaugh and Wentworth 1976), but should not limit use of canonical correlation in situations where correlations along a gradient are of interest. Canonical correlation analysis does not require a multivariate normal distribtuion per se (Morrison 1976), but the closer the data approach normality and linear relationships between variables, the more realistic and more easily interpretable (usually) the relationships between variables will be (Williams 1981). Elsewhere in this volume, Dunn (1981) discusses the use of transformations to arrive at homogeneous variance, an attribute of a multinormal distribution; see also Smith (1977).

Description

Canonical correlation analysis is a technique for finding the correlation between one group of variables, taken as a set, and a second group of variables, also taken as a set, and in many

respects is like principal component analysis with two data sets instead of one. More precisely, we are interested in finding linear combinations of variables in each set that have maximum correlation. Several linear combinations of the two sets of original variables are possible and each pair of functions is so determined as to maximize the new correlation, subject to the restriction that new correlations must be independent of previously defined ones (i.e., orthogonal) (Cooley and Lohnes 1962). The following discussion draws heavily on a readable discussion of canonical correlation by Thorndike (1978). For other useful discussions, see Cooley and Lohnes (1962, 1971), McKeon (1965), Blackith and Reyment (1971), Bock (1975), Harris (1975), and Morrison (1976). Gittens (1979) presents an exhaustive review of the use of canonical correlation in biological sciences, with an emphasis on botanical research.

Definitions

An understanding of the following definitions is important for the discussion of canonical correlation. In all examples addressed in this paper, analyses are concerned with two sets of original variables, such as beach organisms and substrate characteristics, for example. Through the use of iterations, new values are found for the original variables that taken together are maximally correlated. These new values are called canonical variates and the canonical correlation is the maximum Pearson product-moment correlation between variates. It is important to remember that the canonical correlation is not between original variables, but between variates calculated from the variables. Weights used to derive variates from variables are called canonical coefficients and should not be confused with canonical factors which are correlations between the original variables and the variates.

Canonical correlations are invariant statistics since, being product-moment correlations, they are unaffected by linear transformations of the variables in either set. Several authors, e.g., Kettenring (1971), have generalized canonical correlation analysis to three or more sets of original variables, but I know of no biological studies where more than two data sets have been used.

Geometric model

Geometrically, canonical correlation can be considered as a measure of the extent to which entities occupy the same relative position in the space defined by each of the data sets (Cooley and Lohnes 1962). For example, Barkham and Norris (1970) stated that their canonical correlation study attempted to measure the extent to which study sites occupied the same relative position in vegetation space as in soil property space.

A geometric representation of canonical correlation is presented in figure 1. A set of original variables, A and B, defining plane 1 comprise data set 1. Data set 2 is composed of variables C and D and defines plane 2. This example of two sets each with two original variables is the simplest canonical correlation situation; if one data set contained only one variable, this would be a multiple regression problem. Canonical correlation analysis will reveal a composite of A and B, in this example a line called 1_c (fig. 1), that is correlated with composite of C and D, line 2_c. Remembering that the objective is to define new values of the original variables that taken together will be maximally correlated, if the correlation between 1_c and 2_c were the highest that could be found for the data sets 1 and 2, lines 1_c and 2_c would be the first canonical variates. The correlations between A and B with 1_c and C and D with 2_c would be vectors of canonical factors. The canonical correlation (R) is the cosine of the angle between 1_c and 2_c, dotted in figure 1. As the angle approaches 0^o, the canonical correlation approaches 1.00. The canonical correlation is usually squared and R^2 is the proportion of variance in the canonical variate of one set that is accounted for by the variate of the other set. Again, the relationship being described is between the variates, not between the original variables themselves.

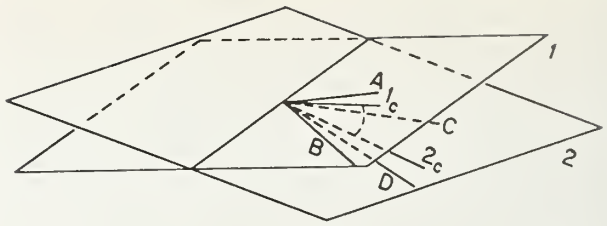


Figure 1. Geometric representation of canonical correlational analysis in two planes A and B. See text for explanation. (from Thorndike 1978)

R^2 is a symmetric index in the sense that the proportion of variance in one set accounted for in the composite of the other set is independent of which set is considered first. In other words, there is no difference in the statistical treatment of the two data sets so that one does not have to make the distinction between a predictor set and criterion or outcome set (Harris 1975).

Test of Significant Roots

As many canonical correlations can be found as there are variables in the smaller set of

original variables. Early workers assumed that only the first canonical correlation was important, although this is not always the case. More than the first canonical correlation may be important depending on the question being asked (Cooley and Lohnes 1962). Bartlett (1941-42) developed a χ^2 test that tests the probability that a particular canonical correlation is significantly different from that which would have been expected with random data. The null hypothesis is that the canonical variates derived from one data set are unrelated to the variates of the second data set. Harris (1976) questioned the use of Bartlett's χ^2 test, but Mendoza et al. (1978) challenged Harris's arguments. Lawley (1959) suggested a slight improvement for Bartlett's χ^2 test of significance.

Thorndike (1978) cautioned that statistical significance does not guarantee meaningful biological relationships, a point which may explain many of the interpretational problems researchers have had with canonical correlation. Also, Bartlett's χ^2 test is very sensitive to departures from normality in data (Barkham and Norris 1970). Blackith and Reyment (1971) comment that biologically meaningful combinations may still be present between data sets that have no statistically significant canonical correlations.

Some Solutions To Problems Of Interpretation

Cross-Validation

Some controversy (e.g., Barcikowski and Stevens 1975) has arisen concerning whether one should investigate relationships between weights (coefficients) or correlations between original variables and variates (canonical factors). Coefficients are usually difficult to interpret, if not meaningless (Cassie and Michael 1968), so that most researchers have investigated canonical factors (e.g., Meredith 1964). As pointed out by Poore and Mobley (1980), examining canonical factors has two advantages: 1) species with high correlations with the same canonical variate are grouped together, and 2) environmental variables which have high correlations with the species groupings are elucidated. Usually one would want to consider any factor above 0.50 (e.g., Webb et al. 1973); factors below 0.30 are probably trivial (Cooley and Lohnes 1971). Bock (1975) has suggested using correlations of the variates from one set with original variables of the <u>other</u> set to characterize between set relationships, but I can find no studies that have applied this technique.

The problem of using correlations between original variables and variates is that interpretation becomes dependent upon the specific relationships between measured variables. Situation-specific covariance, such as that arising from the manner in which data were collected, may become a problem so that one cannot generalize the results to other situations. One would, of course, like the covariance to reflect that of the whole population, not just the subset sampled. Thorndike and Weiss (1973) proposed a check of canonical correlation analysis to discover if any abnormalities in the data are a consequence of site-specific covariance (see also Thorndike 1978). Using this technique, called cross-validation, the original data sets are randomly split into two subsets, canonical correlation is performed on one subset, and relationships in the second subset are examined using the weights derived from the analysis on the first subset. If the results within both subsets are similar, then situation-specific covariance is assumed to be minimal. Also, double cross-validation may be performed whereby both subsets are submitted to canonical correlation analysis and both are cross-checked using weights obtained from the opposite analysis. To my knowledge, no one has attempted this procedure with biological data. One needs a fairly large data set to afford the luxury of splitting the data set in half.

Redundancy

From an interpretational point of view, another major problem with canonical correlation analysis is that R^2 represents variance shared by canonical variates and not variance shared between data sets. Thus, canonical correlation cannot be interpreted as correlations between sets of original variables. A strong correlation may be obtained between two canonical variates even though these variates do not extract significant proportions of variance from the respective sets of original variables. To help with this problem of interpretation, Stewart and Love (1968) developed a measure they called redundancy, whereby one can calculate the amount of variance in one set of original variables that is explained by the variate of the other data set. By calculating redundancy for all variates of a data set and summing the results, the proportion of variance of one set that is accounted for by the other set can be calculated. Since the amount of variance of one set explained by another set does not necessarily equal the amount of variance of the second set explained by the first, the measure of redundancy is not symmetrical (as was R^2) and must be calculated for both data sets. Moreover, the first canonical correlation need not and often does not have the highest redundancy. This procedure is thoroughly discussed in Cooley and Lohnes (1971) and Thorndike (1978).

Several biological investigators have used the measure of redundancy recently and use should increase as the technique becomes more well-known. In an investigation of estuarine diatoms, McIntire (1978) found 41% redundancy in species data, meaning that 41% of variation associated with the

26 taxa used could be accounted for by variation in the six environmental variables measured. Likewise, Poore and Mobley (1980) found that 48% of variation in 22 species of marine benthic animals sampled near a sewage treatment plant outwash was associated with variation in nine environmental variables they measured. Note that in both cases, the reverse redundancy calculation, i.e., how much of the environmental variation can be accounted for by the variation in species, is probably meaningless.

In studying dietary relationships among shrubsteppe passerine birds, Rotenberry (1980) found variation in prey size taken could be accounted for by redundancies of only 24% in horned lark (Eremophila alpestris), 28% in sage sparrow (Amphispiza belli), and 38% in western meadowlark (Sturna neglecta). Grouping species data together, redundancy was still about 30%, leading Rotenberry to conclude that the relationship between diet and morphology is more general in nature than other authors have suggested. One would, of course, have expected high redundancy values if morphology and diet were tightly coupled.

Selection of Variables

All too often investigators cannot interpret the factor correlations that appear between environmental and organismal data. Proper selection of variables may help reduce this problem. Variables should be chosen with some a priori knowledge that a relationship exists between the data sets; the ease with which some environmental variables are measured is no reason to assume that they relate to the distribution of organisms of interest. Also, data should be collected specifically for use in canonical correlation analysis rather than using canonical correlation analysis as an ad hoc or a posteriori approach.

In most environmental studies that have applied canonical correlation analysis, the situation has commonly occurred where investigators have measured many variables, but collected few samples. Since the sample-to-variable ratio must be greater than 1 (the greater the better--see section after next), which variables to exclude from analysis becomes important. Thorndike (1978) found that elimination of some variables in most cases did not change the magnitude of the first canonical correlation. He argued that for a given magnitude, the fewer the variables, the greater the likelihood that a canonical correlation is attributable to real population-wide sources of covariation, rather than to situation-specific covariance.

Poore and Mobley (1980) tried three schemes to reduce the number of variables in their marine benthos study. Initially, they eliminated all rare species (average capture < 1 per sample), reducing the total number of species considered from 246 to 49. For the first analysis, they further eliminated all species with less than 118 individuals sampled which reduced the number of species to 22. They found that this procedure produced results that were very difficult to interpret. For the second analysis they deleted all species found in less than half the sampling stations, reducing the 49 species to 18. This procedure eliminated species found only at a few stations, but since they were interested in the subtle influence of sewage on species distributions, this elimination defeated the purpose of the study. For the third analysis, they deleted species correlated with only 1 or 2 of 13 environmental variables measured and also deleted any environmental variables correlated with 6 or less of the 49 species. This left a total of 22 species and 9 environmental variables. This procedure gave the most easily interpretable results since it eliminated both species and environmental variables which were behaving independently of the pattern they were investigating. Care should be used in this procedure, however, since Cassie (1969) found that adding and subtracting variables at random "capriciously" changed values of the corresponding canonical coefficients. In some situations these changes in coefficient values may be desirable, such as in the empirical step-up and step-down procedures discussed in Thorndike (1978).

Plot of Sample Scores

Interpretation of results of a canonical correlation analysis can usually be aided by plotting individual sample canonical scores in a simple two-dimensional plot. Canonical scores are the summation of canonical variates (data for original variables multiplied by canonical coefficients) and are usually available from any canonical correlation computer program. Standardized scores are used in most cases (Morrison 1976:262). One score will be produced for each sample in each data set. For example, in the study of lizard behavior and microclimate discussed later, James and Porter (1979) plotted sample scores from the first canonical correlation with microclimate on one axis and behavior on the other (see fig. 4). In ecological research, as pointed out by Poore and Mobley (1980), the spatial relationship within the sampling stations (or between data points) may be important.

Sample Size

One goal of canonical correlation analysis is the generalization of results to other situations. In order to reach this goal, one must have faith that canonical coefficients and factors actually reflect underlying relationships; stability in estimates of weights and factors should be expected when sample sizes vary. Monte Carlo examinations of weights and factors (Mendoza et al. 1978) demonstrates that the smaller the sample size, the greater the variability in weight and factor estimates. This argues for large sample

sizes. Thorndike (1978) suggests as a "rule of thumb" that sample size should be the total number of variables in both matrices squared, plus 100 or so for good measure. He suggests minimal sample size to be 10 times the total number of variables plus 100 for good measure.

Few ecological studies have approached sample sizes of the magnitude suggested by Thorndike. Since individual samples in ecological research usually have greater meaning than test results of anonymous individuals in psychological and educational research (Poore and Mobley 1980), a sample size of three or four times the number of variables is probably adequate in most instances. The inherent problem is that as sample sizes approach the number of variables, the canonical correlation approaches 1.00 and statistical significance of the first canonical correlation is assured. Biological significance, however, should be questioned in such situations and results based only on canonical correlation with equal or near equal numbers of samples and variables should be viewed with skepticism. Although it has been suggested (e.g., Green 1971) that multivariate statistics are still useful when the statistical assumptions are relaxed if the results are biologically meaningful (e.g., see Dueser and Shugart 1979), I would still argue for statistical rigor in the application of multivariate statistics.

A case in point is the study by Herrera (1978), in which an attempt was made to depict differences between non-resident and resident passerine bird species in Spain using six measurements of avian community structure to characterize each group. These measurements were calculated for each month and canonical correlation analysis was used to compare monthly changes in community structure. With only 12 samples (1 per month) and 12 variables (6 for each bird group), Herrera (1978) understandably found a high canonical correlation (1.00), resulting in a linear plot with fall, winter and spring avian communities at one extreme and the summer community at the other (fig. 2). A linear ordination is indicative of a 1.00 canonical correlation, and the high canonical correlation is most probably an artifact of the small sample-to-variable ratio. Herrera (1978) discussed at length differences between resident and non-resident canonical factors, but certainly stability of the canonical factors, as well as the lengthy discussion of their importance must be questioned.

Another example of the sample size problem is the investigation of Aart and Smeenk-Enseink (1975) into the distribution of 12 spider species compared to 17 environmental variables. With only 29 samples, this study also found a canonical correlation of nearly 1.00 and the plot of data points was linear.

Several statistical options are now available for examining stability of coefficient and factor estimates. To reduce bias that may be associated with the estimates, especially when sample sizes

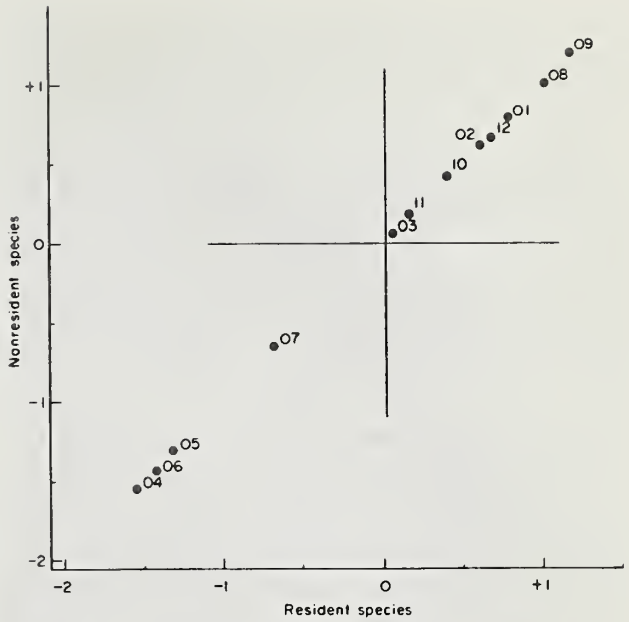


Figure 2. Ordination of 12 monthly samples (e.g., 01 = January) based on canonical correlation analysis (R = 1.00) of resident and non-resident bird communities studied in southeastern Spain. Summer bird communities (April, May, June and July) are separated from other monthly samples. Results are questionable (see text) and linearity of the ordination is due to 1.00 canonical correlation. (from Herrera 1978)

are small, jackknifing estimates is appropriate (see Dempster 1966, Miller 1974). This is a procedure (appendix I) whereby the estimates are recalculated, deleting (usually) one sample from each run. Thus, if one had 15 samples total, 15 canonical correlation runs would be made with 14 samples, each time deleting a different sample. A new value for each estimate is calculated based on the sum of the new runs, and a t-test can determine if the new estimates are significantly different from original ones obtained using all samples (appendix I). If there is no significant difference between original estimates and jackknifed estimates, the confidence one places in the estimates is greatly increased. Confidence in results of the Herrera (1978) and Aart and Smeenk-Enseink (1975) studies would have been greatly improved by jackknifing estimates of canonical factors and coefficients.

Dempster (1966) suggested that for canonical correlation, deletion of one degree of freedom in each run may be superior to deleting complete samples, which have two degrees of freedom. In the example above, 30 new canonical correlations would be run deleting one sample from one of the two data sets each time, and new values for each estimate would be calculated based on the sum of 30 runs.

To reduce the effect of multivariate outliers (data points that do not appear to be members of

the multivariate population under study), canonical correlation using robust estimators can be helpful (Gnanadesikan 1977:127-137). True multivariate outliers can severely affect estimates obtained in most results. Robust estimation decreases, to varying degrees depending on which type of robust estimator is used, the contribution of outlying samples to the estimation of weights (Harner and Whitmore 1981). Gnanadesikan (1977) cautioned that care should be used in the application of robust estimators to small data sets; an apparent "outlier" may in fact only be due to inadequate sampling of the population of interest.

APPLICATIONS OF CANONICAL CORRELATION ANALYSIS

Considering the length of time since development of canonical correlation, the technique has not been applied widely in biological research (see review in Lee [1969]). One of the earliest uses was by Hughes and Lindley (1955) to investigate environmental factors in vegetation trends. Other researchers have attempted to use canonical correlation for phytosociological studies (Austin 1968, Barkham and Norris 1970) but, as mentioned earlier, the use of canonical correlation as an ordinational technique may be limited (Gauch and Wentworth 1976). Webb et al. (1971) found canonical correlation useful in determining the potential of forested areas as agricultural land. Sparling and Williams (1978) suggested the use of canonical correlation for analyzing bird vocalizations, but Martindale (1980) questioned that application.

Correlations Between Morphology and Environment

In animal ecology, some of the most promising applications of canonical correlation have attempted to correlate morphology with some aspect of the environment. In a series of investigations, Jameson and his colleagues used canonical correlation analysis to explore relationships between environmental and morphological variation in the ectothermic Pacific tree frog (Hyla regilla). Since Pacific tree frogs exhibit a wide array of morphological variations which are potentially related to climate and weather through the need to conserve water and energy, canonical correlation analysis was chosen as an appropriate means of depicting the relationships. Calhoon and Jameson (1970) examined relationships between 10 morphological variables and 16 environmental variables that all corresponded to weather patterns the year prior to collection of tree frogs. Tree frogs were collected during one spring from seven different populations in southern California. Discovering that many weather variables were intercorrelated, they performed a second canonical correlation analysis with 8 weather and 10 morphological variables and discussed only the results of the second analysis. Only the first canonical correlation was statistically significant and the results were interpretable as a trade-off between maximizing water conservation by large body size and minimizing heat absorption by small body size. Although not statistically significant, the authors discussed the second canonical correlation as being biologically significant.

In a similar study, Vogt and Jameson (1970) investigated phenotypic variation in one population of Pacific tree frogs sampled monthly over a 3-year period near San Diego, California. In this canonical correlation analysis, 7 morphological measurements of male tree frogs were investigated with 24 environmental (weather) variables covering the 2 years previous to collection of specimens. Five canonical correlations were significant and, taken together, results generally confirm that milder weather correlated with rounder frogs, while decreased rainfall correlated with more elongated animals. Toe length showed the strongest correlation with weather variables suggesting that it responded most rapidly to changes in precipitation. More importantly, canonical correlational analysis elucidated two significant periods in the life of Pacific tree frogs when weather patterns are important: the amount of rain during growth and metamorphosis in spring and the extremes of the high temperatures during fall.

In an expanded version of the two previous investigations, Jameson et al. (1973) examined the relationships of weather patterns, both short- and long-term, to morphological variation in male, female, and juvenile tree frogs from 14 locations in western Oregon. As before, 10 morphological measurements were used and these were correlated to 19 weather variables that characterized weather the year previous to tree frog collection. A second analysis was performed with morphological data using the long-term average (from available weather stations) for the 19 weather variables. The first analysis was termed a "weather" relationship and the second a "climate" relationship. The analyses were performed separately on male, female, and juvenile data sets. The authors concluded that there were meaningful parallels between intraspecific variation and weather and climate. Large males were found in areas of low summer temperatures, but this pattern was not evident in females. Males were more highly correlated with weather patterns, while females were more highly correlated with climatic patterns, a pattern consistent with field observations that males inhabit a more variable microhabitat (ponds). Juvenile morphology was more closely related to weather than climate, presumably because the developmental period they pass through only lasts several weeks. However, the juvenile results were highly variable, suggesting to the authors that there may be selection for variability of response in offspring, a strategy that fecundity of the tree frogs would allow. The major conclusion of this work was that natural selection appears to operate on the size and shape of frogs in different directions during different life history stages and at different localities.

All three tree frog studies illustrate exemplary uses of canonical correlation analysis. Another excellent example is by Boyce (1978) who examined climatic variability and body size in muskrats (Ondatra zibethicus). Results of his analysis are summarized in figure 3. Ten climatic variables were compared with nine morphological measurements and the first canonical correlation was 0.776. (The first six canonical correlations were significant.) Body length (BODY), total length (TOT), condylobasal length (CB) and zygomatic breadth (ZB) were all highly correlated with the first morphological variate and represent body size variation. This is not surprising since the first canonical correlation will almost always be a size-related axis. Longitude (LONG), annual range of evapotranspiration (R_{ev}), and annual coefficient of variation of monthly evapotranspiration (S_{ev}) had high correlations with the climatic variate. Boyce interpreted these results as showing that body size variation is positively correlated with climatic seasonality. Discovering this relationship supported the notion that herbivore body size is positively correlated with primary productivity. Boyce concluded (in part) that the larger individuals may be favored in seasonal environments because they can subsist longer without food.

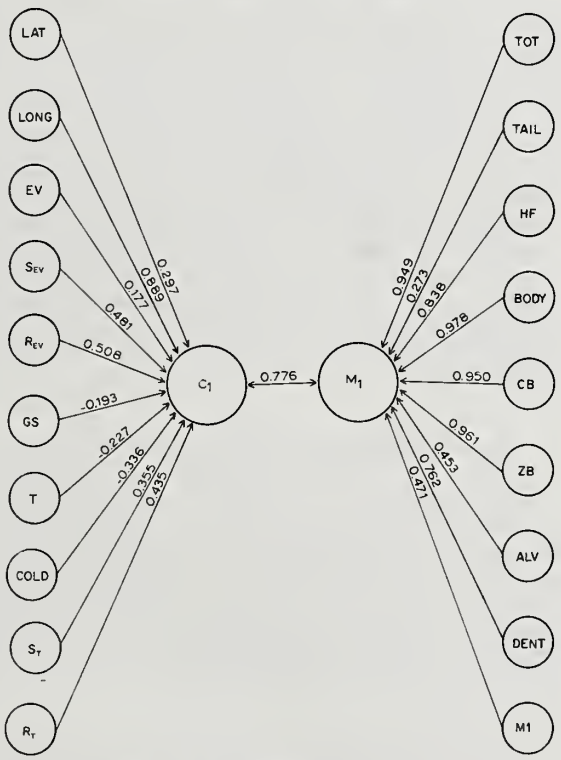


Figure 3. Diagram of the first pair of canonical variates in study of climatic variability and body size variation in muskrats (from Boyce 1978). Definitions of important climatic variables and morphological variables are given in text. C_1 is the first climatic canonical variate, M_1 is the first morphological canonical variate, 0.776 is the canonical correlation, and the numbers along the arrows are correlations between original variables and canonical variate.

Another important study is that of Karr and James (1975) in which 196 species of birds from several continents were compared using 17 morphological measurements and 14 ecological variables, asking the question "What are the ecological correlates of patterns of morphological variation?" The first six canonical correlations were statistically significant and relationships between morphology and ecology became apparent: relatively longer thinner bills and smaller body size (first correlation), relatively longer legs and narrower bills with ground foraging (second correlation), etc. (see Karr and James [1975:table 6] for complete listing). The authors chose to discuss in depth the second canonical correlation since the first correlation related primarily to size phenomena, the results of which were somewhat misleading due to the influcence of the morphologically extreme hummingbirds and sunbirds.

Karr and James (1975) cautioned that care should be used in the selection of variables and generalization of results. They purposely restricted analysis to primarily forest-inhabiting bird species and did not generalize their results beyond that. Addition of, say, water-dependent species such as ducks and herons would have changed the results dramatically (due to the changes in morphology) and the scope of the study would have changed accordingly. Occasionally researchers have combined variables which do not appear to have any apparent relation. For example, DesGrandes (1978) related seven morphological measurements of 21 hummingbird species in Mexico with five ecological variables in order to support his field observations that large birds tend to be sedentary and dominant while smaller species tend to be highly vagile and subordinate. The canonical factors from the first canonical correlation that demonstrate this relationship are presented in table 1. It is unclear what relationship "altitude" has with other variables. Such variables should be culled from canonical correlation analyses and only variables that clearly relate to the same phenomenon should be included in the respective data sets.

Behavioral Applications

A recent paper by James and Porter (1979) demonstrated the applicability of canonical correlation to ecological behavioral investigations. They were interested in relating behavior of the African rainbow lizard (Agama agama) to microclimatic variables. Six microclimatic variables were measured at hourly intervals over several days in October, January, and April and the behavior of a dominant male lizard was also categorized at the same hourly intervals. The authors used canonical correlation

Table 1. Canonical factors associated with first canonical root (R = 0.93) showing relationship between dominance and large body size of hummingbirds contrasted with subordinance and small body size. Data from analysis of Mexican hummingbird communities (DesGrande[3]).

Ecological variables	Factors	Morphological variables	Factors
Dominance	0.92	Length	0.85
Altitude	-0.08	Culmen	0.71
Status	-0.65	Bill width	0.85
Territory	0.40	Bill depth	0.89
Food	0.19	Bill curve	0.14
		Wing disc loading	-0.23
		Dimorphism	0.65

[3]Personal communication with J-L. DesGrandes, Canadian Wildlife Service, Quebec Region, Ste-Foy, Quebec, Canada.

analysis to investigate relationships between behavior and microclimate, and obtained a nice depiction of the effect of heat load on behavior (fig. 4). The distances between columns is a function of the relationship between responses to low (body flattened) and high (body vertical) heat loads. Obtaining the same pattern in October and April argued against a seasonal variation in these relationships.

This work is also important because it employs two methods which improve the usefulness of a canonical correlation analysis. First, although only six microclimatic variables were actually measured in the field, 17 were used in the analysis. These additional 11 variables were constructed by defining various interactions between variables (see James and Porter [1979] for details). The technique of building interactions into the model is used widely in multiple regression, but few workers have considered applying it to canonical correlation analysis.

The second improvement concerns the removal of singularities. Since behavioral data are mutually exclusive, i.e., a lizard cannot be doing two behaviors at the same time, all behavioral rows add up to the same value (1 since all data were categorized). Therefore, the behavior "resting" was eliminated from the analysis so that the sum of the rows was now different. The elimination of one column of data does not affect the conclusions, and in this case, removed the singularity.

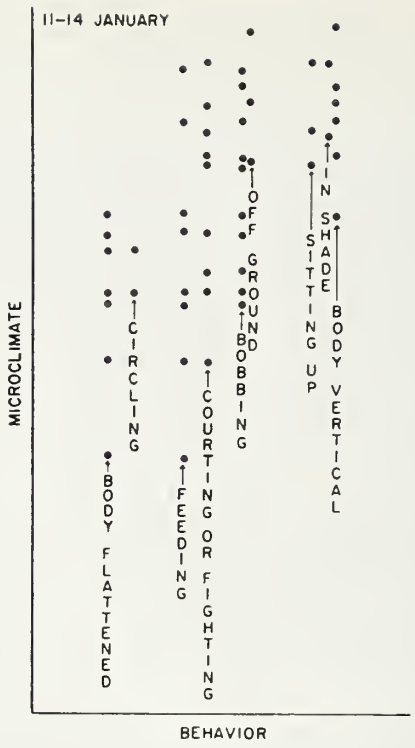


Figure 4. Representation of the relationship between microclimate and behavior of the African rainbow lizard based on canonical correlation analysis (R = 0.72) of data collected between 11-14 January 1971 in Ghana (from James and Porter 1979). Ordination depicts behavioral changes as heat load increases. Each point represents an hourly measurement of behavior and microclimate.

Animal Community Ecological Applications

Using canonical correlation analysis, Webb et al. (1973) were able to present a land-use scheme for nature conservation in Australia based on vegetational characteristics and bird communities of tropical forests. Few other studies have used canonical correlation analysis in terrestrial community ecology, and most that have, suffered from small sample-to-variable ratios (e.g., Herrera 1978).

Use of canonical correlation analysis is increasing in marine ecological studies, which is understandable since specific environmental factors enter prominently into the distribution of many marine invertebrates. One of the first applications was by Cassie and Michael (1968) in a study of invertebrate distribution and sediment size on an intertidal mud flat near Auckland, New Zealand. They were interested in the relationship between eight invertebrate species and nine sediment variables sampled at 21 locations. The analysis suffered from a small sample to variable ratio, and the authors obtained much better results with principal component analysis for

which they had 40 samples. But, realizing the sample-to-variable ratio problem, they suggested that canonical correlation analysis may still be useful in larger studies.

Two recent studies have been more successful with canonical correlation analysis. McIntire (1978) investigated distribution of 26 non-planktonic diatom taxa based on six environmental variables in an estuary on the Oregon coast. The first canonical correlation (R=0.98) ordinated the species along a salinity gradient. The second canonical correlation (R=0.95) highlighted mean daily salinity range and mean temperature, which McIntire interpreted as a gradient to salt tolerance (stenohaline to euryhaline species). In Victoria, Australia, Poore and Mobly (1980) used canonical correlation analysis to ordinate 22 benthic invertebrate species based on nine environmental variables measured at 36 stations at the outwash of a sewage treatment plant. Although the sample-to-variable ratio was small, the authors still found useful results. The first canonical correlation (R=0.99) reflected a depth gradient based primarily on amount of fine sand associated with shallow waters. The second canonical correlation (R=0.98) separated the middle samples from the shallow and deep water samples (figure 5). Such a "horseshoe-shaped" ordination is to be expected when attempting to ordinate species distributions along a steep gradient (e.g., see Phillips [1978]).

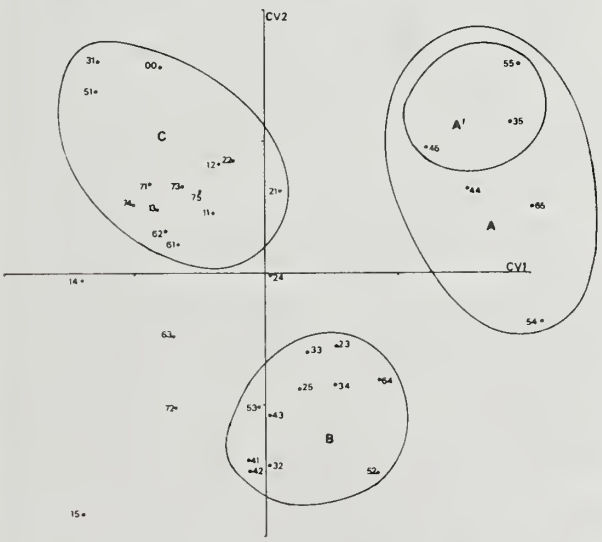


Figure 5. Ordination of benthic marine samples from outwash of sewage treatment plant based on first two canonical variates (CV1 and CV2) (from Poore and Mobley 1980). Separation is along depth gradient: A' = very deep samples, A = deep samples, B = intermediate depths, and C = shallow water samples (ellipses drawn by eye). Placement of some sample stations outside of ellipses (e.g., 14, 72, and 63) is due to unusual environmental conditions at those sites (see Poore and Mobley 1980 for details).

CONCLUSIONS

The major point that researchers who use canonical correlation should keep in mind is that the larger the sample-to-variable ratio, the more stable canonical factors and weights become, so that sample sizes should be as large as possible. As relationships between variables approach linearity and data sets approach a multivariate normal distribution, results should become more easily interpretable.

By perusing the works that have successfully used canonical correlation to date, it becomes clear that experimental design and planning are critical. It is imperative that data be collected with canonical correlation in mind. All too often application of multivariate statistics appears to be an ad hoc or a posteriori approach rather than the a priori approach that it should be. The selection of variables is extremely important and much of the past troubles with interpretation may relate to researchers having no idea how the environmental variables measured actually influence the organism(s) of interest. The problem of adequate sample size may make use of the technique prohibitive, although this has rarely stopped anyone yet!

With proper planning, canonical correlation analysis can become a very useful tool for wildlife habitat studies, and probably will become more widely used as researchers become familiar with it. Planning is certainly the key to successful use and cannot be stressed too heavily.

ACKNOWLEDGMENTS

Frances James kindly read an early draft of this paper and made many helpful suggestions concerning terminology and explanations of examples used in the text. Through her creative use of multivariate statistics, including canonical correlation analysis, she has made great contributions in ecological research and continues to be a leader in the field of applying multivariate statistics to ecological problems.

Mary Barkworth pointed out many inconsistencies in an earlier draft. Other comments were made by Douglas Anderson, James A. MacMahon, Donald Phillips, David Schimpf, Jeffrey J. Short, and Eric Zurcher. While writing this paper I was supported by grants DEB 78-05328 (NSF) to James A. MacMahon and 03-5-022-84-NOAA to Robert W. Riseborough.

LITERATURE CITED

Aart, P.J.M. van der, and N. Smeenk-Enserink. 1975. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherland. Journal of Zoology 25:1-45.

Austin, M.P. 1968. An ordination study of a chalk grassland community. Journal of Ecology 56:739-757.

Barcikowski, R.S.,and J.P. Stevens. 1975. A Monte Carlo study of the stability of canonical correlations, canonical weights and canonical variate-variable correlations. Multivariate Behavorial Research 10:353-364.

Barkham, J.P., and J.M. Norris. 1970. Multivariate procedures in an investigation of vegetation and soil relations of two beech woodlands, Cotswold Hills, England. Ecology 51:630-639.

Bartlett, M.S. 1941-42. The statistical significance of canonical correlations. Biometrika 32:29-37.

Blackith, R.E., and R.A. Reyment. 1971. Multivariate morphometrics. 412p. Academic Press, New York, N.Y.

Bock, R.D. 1975. Multivariate statistical methods in behavioral research. 623p. McGraw-Hill, New York, N.Y.

Boyce, M.S. 1978. Climatic variability and body size variation in the muskrats (Ondatra zibethicus) of North America. Oecologia 36:1-19.

Bryant, E.H., and W.R. Atchley. 1975. Multivariate statistical methods, within-groups covariation. Benchmark papers in systematic and evolutionary biology 2. 436p. Dowden, Hutchinson and Ross, Stroudsburg, Penn.

Calhoon, R.E., and D.L. Jameson. 1970. Canonical correlation between variation in weather and variation in size in the Pacific tree frog, Hyla regilla, in southern California. Copeia 1970:124-134.

Cassie, R.M. 1969. Multivariate analysis in ecology. Proceedings of New Zealand Ecological Society. 16:53-57.

Cassie, R.M., and A.D. Michael. 1968. Fauna and sediments of an intertidal mudflat: A multivariate analysis. Journal of Experimental Marine Biology and Ecology 2:1-23.

Cooley, W.W., and P.R. Lohnes. 1962. Multivariate procedures for the behavioral sciences. 211p. John Wiley and Sons, New York, N.Y.

Cooley, W.W., and P.R. Lohnes. 1971. Multivariate data analysis. 364p. John Wiley and Sons, New York, N.Y.

Dempster, A.P. 1966. Estimation in multivariate analysis. p. 315-334. In Krishnaiah, P.R.,editor. Multivariate analysis. 592 p. Academic Press, New York, N.Y.

DesGrandes, J-L. 1978. Organization of a tropical nectar feeding bird guild in a variable environment. Living Bird 17:199-236.

Dueser, R.D., and H.H. Shugart, Jr. 1979. Niche pattern in a forest floor small-mammal fauna. Ecology 60:108-118.

Dunn, J.E. 1972. Discussion. p. 128-130. In Allen R.T., and F.C. James, editors. A symposium on ecosystematics. University of Arkansas Museum Occasional Papers No. 4, Fayetteville, Ark.

Dunn, J.E. 1981. Data-based transformations in multivariate analysis. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Gauch, H.G., Jr., and T.R. Wentworth. 1976. Canonical correlation analysis as an ordination technique. Vegetatio 33:17-12.

Gittens, R. 1979. Ecological applications of canonical analysis. p. 309-535. In Orloci, L., C.R. Rao, and W.M. Stiteler, editors. Multivariate methods in ecological work. Statistical ecology series, volume 7. 550p. International Co-operative Publishing House, Fairland, Md.

Gnanadesikan, R. 1977. Methods of statistical data analysis of multivariate observations. 311p. John Wiley and Sons, New York, N.Y.

Green, R.H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of central Canada. Ecology 52:593-556.

Harner, E.J., and R.C. Whitmore. 1981. Robust principal component and discriminant analysis of two grassland bird species' habitat. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [April 23-25, 1980, Burlington, Vt.]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Harris, R.J. 1975. A primer of multivariate statistics. 332p. Academic Press, New York, N.Y.

Harris, R.J. 1976. The invalidity of partitioned-U tests in canonical correlation and multivariate analysis of variance. Multivariate Behavioral Research 11:353-365.

Herrera, C.M. 1978. Ecological correlates of residence and non-residence in a Mediterranean passerine bird community. Journal of Animal Ecology 47:871-890.

Hotelling, H. 1935. The most predictable criterion. Journal of Educational Psychology 26:139-142.

Hotelling, H. 1936. Relations between two sets of variates. Biometrika 28:321-377.

Hughes, R.E., and D.V. Lindley. 1955. Application of biometric methods to problems of classification in ecology. Nature 175:806-807.

James, F.C., and W.P. Porter. 1979. Behavior-microclimatic relationships in the African rainbow lizard, Agama agama. Copeia 1979: 585-593.

Jameson, D.L., J.P. Mackey, and M. Anderson. 1973. Weather, climate, and the external morphology of Pacific tree toads. Evolution 27:285-302.

Karr, J.R., and F.C. James. 1975. Ecomorphological configurations and convergent evolution. p. 258-291. In Cody, M.L., and J.M. Diamond, editors. Ecology and evolution of communities. 545p. Belknap Press of Harvard University, Cambridge, Mass.

Kettenring, J.R. 1971. Canonical analysis of several sets of variables. Biometrika 58:433-451.

Lawley, D.N. 1959. Tests of significance in canonical analysis. Biometrika 46:59-66.

Lee, P.J. 1969. The theory and application of canonical trend surfaces. Journal of Geology 77:303-318.

Martindale, S. 1980. On the multivariate analysis of avian vocalizations. Journal of Theoretical Biology 83:107-110.

McIntire, C.D. 1978. The distribution of estuarine diatoms along environmental gradients: A canonical correlation. Estuarine and Coastal Marine Science 6:447-457.

McKeon, J.J. 1965. Canonical analysis: Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory. Psychometric Monographs 13.

Mendoza, J.L., V.H. Markos, and R. Gonter. 1978. A new perspective on sequential testing procedures in canonical analysis: A Monte Carlo evaluation. Multivariate Behavioral Research 13:371-382.

Meredith, W. 1964. Canonical correlations with fallible data. Psychometrika 29:55-65.

Miller, R.G. 1974. The jackknife - a review. Biometrika 61:1-15.

Morrison, D.F. 1976. Multivariate statistical methods. Second edition. 415p. McGraw-Hill Book, New York, N.Y.

Phillips, D.L. 1978. Polynomial ordination: Field and computer simulation testing of a new method. Vegetatio 37:129-140.

Pielou, E.C. 1977. Mathematical ecology. 385p. John Wiley and Sons, New York, N.Y.

Poore, G.C.B., and M.C. Mobley. 1980. Canonical correlation analysis of marine macrobenthos survey data. Journal of Experimental Marine Biology and Ecology 45:37-50.

Rao, C.R. 1955. Estimation and tests of significance in factor analysis. Psychometrika 20:93-111.

Rotenberry, J.T. 1980. Dietary relationships among shrubsteppe passerine birds: competition or opportunism in a variable environment? Ecological Monographs 50:93-110.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58:810-819.

Sparling, D.W., and J.D. Williams. 1978. Multivariate analysis of avian vocalizations. Journal of Theoretical Biology 74:83-107.

Stewart, D., and W. Love. 1968. A general canonical correlation index. Psychological Bulletin 70:160-163.

Thorndike, R.M., 1978. Correlational procedures for research. 340p. Gardner Press, New York, N.Y.

Thorndike, R.M., and D.J. Weiss. 1973. A study of the stability of canonical correlations and canonical components. Educational and Psychological Measurement 33:123-134.

Vogt, T., and D.L. Jameson. 1970. Chronological correlation between change in weather and change in morphology of the Pacific tree frog in southern California. Copeia 1970:135-144.

Webb, L.J., J.G. Tracy, W.T. Williams, and G.N. Lance. 1971. Prediction of agricultural potential from intact forest vegetation. Journal of Applied Ecology 8:99-121.

Webb, L.J., J.G. Tracey, J. Kikkawa, and W.T. Williams. 1973. Techniques for selecting and allocating land for nature conservation in Australia. p. 39-52. In Costin A.B., and R.H. Groves, editors. Nature conservation in the Pacific. 337 p. Australian National University Press, Canberra, Australia.

Williams, B.K. 1981. Discriminant analysis in wildlife research: theory and applications. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [April 23-25, 1980, Burlington, Vt.]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

APPENDIX I

Jackknife estimation

(written by James E. Dunn)

$\hat{\theta}$ = usual estimate based on all data

Divide data set at random into g groups of h observations each, e.g., typically g = n, h = 1.

Compute $\hat{\theta}_{-i}$, i.e., usual estimate with ith group deleted, i=1,2,...,g

Compute pseudo values $\tilde{\theta}_i = g\hat{\theta} - (g - 1)\,\hat{\theta}_{-i}$

i = 1,2,..., g

Jackknifed estimate is

$$\tilde{\theta} = \frac{1}{g} \sum_{i=1}^{g} \tilde{\theta}_i$$

$$s^2_{\tilde{\theta}} = \widehat{VAR[\tilde{\theta}]} = \sum_{i=1}^{g} \frac{(\tilde{\theta}_i - \tilde{\theta})^2}{g-1}$$

Test of significance: $H_o : \tilde{\theta} = \theta_o$

$$t = \frac{\tilde{\theta} - \theta_o}{s_{\tilde{\theta}}/\sqrt{g}} \sim t_{(g-1)}$$

DISCUSSION

JAKE RICE: In your talk you stressed that canonical correlation was exactly that: a correlation method, and one does not need to identify predictor and criterion sets of variables. Suppose you did have some biological reason for identifying one set of variables as predictors and the other set as criteria variables (as is common in simple regression studies); would (or could) one do anything different that might improve the findings? (As in some contexts regression methods provide results preferable to correlation methods on the same data set.)

KIMBERLY SMITH: James Dunn has an answer to that question.

JAMES DUNN: If SAS PROC GLM is available, write a MODEL statement with the criteria variables on the left and the predictor variables on the right. MANOVA mode will work equally well if the predictors are continuous or categorical. In the former, it may be necessary to arrange the MODEL statement properly, specify Type I SS, and accumulate the derived number of H matrices into a composite H_c before computing a likelihood ratio test based on

$$\ln\left(\frac{|E|}{|E + H_c|}\right)$$

or a union-intersection test based on max $ch(E^{-1}H_c)$.

JOSEPH FOLSE: In multiple regression analysis, inclusion of extraneous variables does not bias other regression estimates whereas deletion of important variables does. What effects does this have in canonical correlations?

KIMBERLY SMITH: One way that has been suggested of attempting to improve canonical correlations is to eliminate variables that have low loadings (or weights) and repeat the analysis with the reduced model. Interestingly, Cassie (1969) has suggested, however, that if one starts to delete variables randomly, the relationships obtained from canonical correlation change rather capriciously. Thus, the same relationship appears to hold true for canonical correlation - inclusion of extraneous variables should not bias other correlation estimates, whereas deletion of important variables will.

DATA-BASED TRANSFORMATIONS IN MULTIVARIATE ANALYSIS[1]

James E. Dunn[2]

Abstract.--Univariate transformations are considered initially, because of the common practice of transforming separately the marginal distribution of each variable of a multivariate observation. Familiar examples include those based on a priori assumptions about the underlying sampling distribution, as well as several general classes of empirical transformations recommended in a recent text by Mosteller and Tukey. Multi-normal criteria are considered as a basis for obtaining and evaluating multivariate transformations, including the likelihood criterion and various transformations to uniform statistics. The extension of power and shifted-power transformations to multivariate analysis is reviewed in detail, including recently published work involving q-sample problems.

Key words: Multivariate; power transformation; uniform statistics.

INTRODUCTION

At first exposure, a student without data typically will be intimidated by the wilderness which exists in the world of data transformations. "How," he will ask, "can I interpret a significant difference between means of logarithms of my data?" With data he will seem perfectly comfortable in comparing mean pH values ($-\log H^+$). When life testing with an automatic termination date, say, after 15 days, he will automatically compare means of reciprocal response times, where 'starter' values of 1/16 are routinely assigned for any survivors. An arbitrary error of at most 1/16 for any survivor seems perfectly acceptable in this context. If asked to compare the means when sampling from two populations whose variances clearly are unequal, he will agree that any comparison is ambiguous until a common metric (σ) is established through use of a variance stabilizing transformation.

Even though this manuscript is directed toward multivariate problems, univariate transformations are considered initially because of the common practice of transforming the marginal distribution of each variable separately. Even though marginal normality does not guarantee multivariate normality, as may be demonstrated in rather contrived examples, numerous cases have occurred where this approach has seemed sufficient.

Multivariate transformations are introduced in the following section in the natural language of matrix notation. Capital letters are used to denote matrices (e.g., A, I, Σ), lower case letters underscored for column vectors (e.g., $\underline{y}$, $\underline{j}$, $\underline{\beta}$), and lower case letters in general for scalars, either random variables or constants, including the elements of vectors and matrices.

A natural danger in a paper of this type is to focus on the taxonomy rather than the systematics (ecology) of transformations. To avoid this, I have tried to emphasize what I consider in the process of selecting a data-based transformation.

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Professor and Statistical Consultant, Department of Mathematics and Statistics, University of Arkansas, Fayetteville, AR 72701.

In the interest of space, two interesting categorical response situations can only be mentioned here. The GSK model (Grizzle, Starmer, and Koch 1969) for analysis of categorical response using linear models has promise for studying multivariate relative frequency problems, such as bird counts by species at different locations, or stomach contents by taxonomic group. Dunn and Cappy (1979a, 1979b) have studied the problem of invertible transformations in this context. The multivariate logistic model $p = [1 + \exp(-\beta'\underline{x})]^{-1}$ has come into focus as a robust classifier compared to multi-normal classification functions (cf. Efron 1975, Press and Wilson 1978). In particular, it lends itself to use of categorical response variables and non-linear boundaries.

UNIVARIATE TRANSFORMATIONS

Even though the material in this section is generally known, it is reviewed here for completeness. If $y \sim N(\mu, \sigma^2)$, then it is evident from the uniqueness of the cumulant generating function $K_y(t) = \mu t + \sigma^2 t/2$ that the normal distribution is the only possible distribution in which the mean and variance are functionally unrelated. Since most of our familiar hypotheses tests (e.g., t, F, χ^2, etc.) depend on the normality assumption, finding a transformation of non-normal data whose variance is unrelated to the mean seems to be a logical approach to normalization.

Variance Stabilization From A Priori Assumptions

Suppose that y has mean μ and variance $\sigma^2(\mu)$, and we require a transformation $t = t(y)$ such that var[t] is a constant c to some acceptable order of approximation. Using the familiar stochastic Taylor's series expansion of t(y) about μ leads to the well-known defining equation (cf. Rao 1952)

$$t(\mu) = \int \sqrt{c/\sigma^2(\mu)}\, d\mu \quad . \qquad (1)$$

For example, suppose that y has a Poisson distribution with $\mu = \lambda$ and $\sigma^2(\mu) = \lambda$.

Then

$$t(\lambda) = \sqrt{c} \int \lambda^{-1/2}\, d\lambda = 2\sqrt{c}\,\sqrt{\lambda} = \sqrt{\lambda} \ ,$$

provided we specify the approximate variance to be $c = 1/4$. In actual fact if $t = \sqrt{y}$, then $var[t] = 1/4 + 3/(32\lambda) + (\cdot)/\lambda^2 + \ldots = 1/4 + O(1/\lambda)$. Anscombe (1948) produced a still more stable transformation $t = \sqrt{y + 3/8}$ with $var[t] = 1/4 + O(1/\lambda^2)$. Clearly, variance stabilization is more effective for Poisson counts with large rather than small means. Familiar transformations, including improvements suggested by Anscombe (1948) appear in table 1. They are data-based in the sense that they reflect our prior beliefs about the parent population being sampled.

A Data-Based Power Transformation

Suppose that y is sampled from a population with mean μ and $\sigma^2(\mu) = \alpha\mu^b$, i.e the variance is proportional to some unknown power of the mean. Using equation (1), one obtains the result

$$t(\mu) = \sqrt{c/\alpha} \int \mu^{-\beta/2}\, d\mu$$

$$= \begin{cases} \mu^{1-\beta/2} \text{ with } c = \alpha(1 - \beta/2)^2 \text{ if } \beta \neq 2 \\ \ell n(\mu) \text{ with } c = \alpha \text{ if } \beta = 2. \end{cases}$$

If a sample is available from each of q populations, or if a large sample from a single population is divided at random into q subsamples, then one may estimate β from the sample means and variances $\{(\bar{y}_i, s_i^2): i=1,\ldots,q\}$ as the slope of the regression of $\ell n\ (s_i^2)$ on $\ell n\ (\bar{y}_i)$. Then $t = \ell n(y)$ or $t = y^{1-\hat{\beta}/2}$, depending on the proximity of $\hat{\beta}$ to 2.

Mosteller-Tukey Transformations

Mosteller and Tukey (1977) approach the problem of selecting a transformation using only prior knowledge that certain parts of the range of t(y) should represent a differential shrinkage or expansion of the domain of y. Selecting a proper transformation is visualized to be an evolutionary process, i.e., try, look, and try again, with interactive access to the data set assumed. Visual display, to look for linear trend or parallel trends, is important to see if progress is being made in the sequence of steps. General families of transformations suggested by Mosteller and Tukey (1977) are summarized in terms of the scale of measurement required. Most readers will feel at ease with their use of 'starter' values t(y + c), though justification of the recommended c = 1/6 is somewhat mysterious.

Amounts and counts. Assuming that $y \geq 0$, one may use

$$t(y) = \begin{cases} A(y/A)^p/p + (1 - 1/p)A \text{ if } p \neq 0 \\ A\ \ell n(y/A) + A \text{ if } p = 0 \end{cases}$$

if a match t(y) = y is required when y = A. A choice of $p > 1$ will cause expansion of the tails of the t axis, with the converse holding for $p < 1$. If the data contain zeros, then use of t(y + 1/6) is suggested.

Table 1. Familiar univariate, variance-stabilizing transformations.

Distribution/Mean/Variance	t	var[t]
Poisson/λ/λ	$\sqrt{y + 3/8}$	$1/4 + 0(1/\lambda^2)$
Binomial/np/np(1-p)	$\sin^{-1}\sqrt{(y + 3/8)/(n + 3/4)}$	$(4n + 2)^{-1}$
	where n must be a constant.	
Negative binomial/ m/m(1 + m/k)	$\sinh^{-1}\sqrt{(y + 3/8)/(k - 3/4)}$, $k \geq 2$	$\psi(k)/4$
	$\ell n(y + k/2)$, $k < 2$	$\psi(k)$
	where k must be known.	
s^2 (Normal population)/ $\sigma^2/(2\sigma^2/(n-1))$	$\ell n(s^2)$	$2/(n-1)$
	where n must be a constant.	

Counted fractions. The typical situation is "y successes in n trials". If a started fraction is defined to be

$$f = (y + 1/6)/[y + 1/6) + (n - y + 1/6)]$$

$$= (y + 1/6)/(n + 1/3),$$

then either the folded square root (froot)

$$t(y) = \sqrt{2}(\sqrt{f} - \sqrt{1 - f})$$ or the folded logarithm (flog)

$$t(y) = 1/2\ \ell n\ f - 1/2\ \ell n(1-f) = \ell n\sqrt{f} - \ell n\sqrt{1-f}$$

$$= 1/2\ \ell n(y + 1/6) - 1/2\ \ell n(n - y + 1/6)$$

is recommended. Note that f = 1/2 and t = 0 in either case when y = n/2. Hence, a 50% response corresponds to a symmetry point for t, with little differential expansion in the neighborhood of t = 0 and considerable expansion in either tail.

Ranks. If n objects are assigned ranks $r_1,\ldots,r_n$, then the counted fraction transformation may be applied as

$$t_i = 1/2\ \ell n[(r_i - 1/3)/(n + 1 - r_i - 1/3)]$$

$$(i=1,\ldots,n).$$

Alternatively, 'rankits' variously defined as

$$t_i = \begin{cases} \Phi^{-1}\ [(r_i - 1/3)/(n + 1/3)] & \text{(Tukey)} \\ \Phi^{-1}\ [(r_i - 3/8)/(n + 1/4)] & \text{(Blom)} \\ \Phi^{-1}\ [(r_i/(n + 1)] & \text{(Vander Waerden)} \end{cases} (i=1,\ldots,q)$$

are also commonly used in this context (cf. SAS 1979), where Φ is the standard normal cumulative distribution function (c.d.f.).

Grades. Here, the observations are simply classifed into one of q mutually exclusive, but ordered classes $a_1, a_2,\ldots,a_q$. Suppose that the observations are to be scored in terms of an assumed standard distribution with c.d.f. F(y). If the relative class frequencies are $p_1, p_2,\ldots,p_q$ ($\sum_{j=1}^{q} p_j = 1$), and $s_i = \sum_{j=1}^{i} p_j$, then cut-points are defined by $c_0 = -\infty$, $c_q = \infty$ and $c_i = F^{-1}(s_i)$ for i=1,...,q-1. The score μ_i to be assigned to an observation in the i'th class is the conditional mean of the standard distribtuion on the interval (c_{i-1},c_i), that is

$$\mu_i = \int_{c_{i-1}}^{c_i} y\ dF(y)/p_i \quad (i=1,\ldots,q).$$

In the normal case with $F(\cdot) = \Phi(\cdot)$, this becomes

$$\mu_i = [\exp(-c_{i-1}^2/2) - \exp(-c_i^2/2)]/p_i\sqrt{2\pi},$$

(cf. Kendall and Stuart 1967), while for the logistic as the standard distribution, Mosteller and Tukey (1977) give

$$\mu_i = [h(s_i) - h(s_{i-1})]/p_i,$$

where $h(s) = s\ \ell n(s) + (1 - s)\ell n(1 - s)$. Note the attractive feature that $c_1,\ldots,c_q$ need not be obtained explicitly in the latter case. In any case, considerable time generally will be required to complete the data recoding in multivariate applications. Cockrell (1980) will distribute (on request) a SAS procedure RECODE which seems to fit this application quite nicely.

MULTIVARIATE TRANSFORMATIONS

In a search for a class of generally useful transformations of a univariate response y, Box and Cox (1964) proposed the power transformation

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \text{for } \lambda \neq 0, \\ \ell n(y) & \text{for } \lambda = 0, \end{cases} \tag{2}$$

provided $y > 0$, or the shifted-power transformation

$$y^{(\lambda)} = \begin{cases} [(y + \xi)^{\lambda} - 1]/\lambda & \text{for } \lambda \neq 0 \\ \ell n(y + \xi) & \text{for } \lambda = 0, \end{cases} \tag{3}$$

for $y + \xi > 0$, where ξ either may be a starter value supplied by the user or else treated as an additional parameter to be estimated. Since typical statistics such as t and F are invariant with respect to changes in location and scale, use of $y^{(\lambda)}$ is identical to use of y^{λ} or $(y + \xi)^{\lambda}$, thus equivalent to the power transformation derived above. The particular form of $y^{(\lambda)}$ simply emphasizes the fact that $\ell n(y)$ is a continuity point of $y^{(\lambda)}$ at $\lambda = 0$, that is

$$\lim_{\lambda \to 0} y^{(\lambda)} = \lim_{\lambda \to 0} (y + \xi)^{\lambda}\, \ell n(y + \xi) = \ell n(y + \xi).$$

If $\underline{y}' = (y_1, \ldots y_p)$ is a p-variate response vector, then Andrews et al. (1971) suggested a multivariate extension by defining $\underline{y}^{(\lambda)} = [y_1^{(\lambda_1)}, \ldots, y_p^{(\lambda_p)}]'$, where $y_k^{(\lambda_k)}$ is defined by equations (2) or (3) in terms of parameter sets $\{\lambda_k\}$ or $\{\lambda_k, \xi_k\}$ for $k = 1, \ldots, p$.

In order to choose 'best' values for either of the above parameter sets in any particular application, we must specify our criteria. First, since almost all multivariate analyses depend on statistics which are functions of the covariance matrix, (e.g., principal components, canonical correlation, discriminant functions, and multivariate classification), we shall want a transformation to new variables whose covariance matrix contains all of the information about inter-relationships among the variables (independently of the mean). Again, considering the uniqueness of the cumulant generating function $K_{\underline{y}}(\underline{t}) = \underline{\mu}'\underline{t} + \underline{t}'\Sigma\underline{t}/2$ for $\underline{y} \sim N(\underline{\mu}, \Sigma)$, this suggests that any criteria for the transformation be based on a characterization of multi-normality. Two of these possibilities are considered in the remainder of this section. Second, since the development of a transformation is likely to be an iterative process, we shall require an easily comprehended measure of our progress toward the optimum and an assessment of our final result. Finally, since in most applications, the original scales of measurement are meaningful, only coordinate dependent approaches will be considered here [cf. Cox and Small (1978) for coordinate independent approaches].

Likelihood Criterion

Let us consider the general q-sample ($q \geq 1$), multivariate case [which includes the q-sample, univariate case discussed by Box and Cox (1964), and the one-sample, multivariate case treated by Andrews et al. (1971)]. Suppose that a random sample $\underline{y}_{i1}, \ldots, \underline{y}_{i,n_i}$ of size n_i is drawn independently from each of q, p-variate populations ($i=1,\ldots,q$). Considering the simpler case of the transformations defined by equation (2), we shall want to find an estimate of the parameter set $\{\lambda_k\}$ such that

$$\underline{y}_{ij}^{(\lambda)} \sim N(\underline{\mu}_i, \Sigma), \text{ where} \tag{4}$$
$$(\underline{y}_{ij}^{(\lambda)})_k = (y_{ijk}^{\lambda_k} - 1)/\lambda_k$$

for $i=1,\ldots,q$; $j=1,\ldots,n_i$; and $k=1,\ldots,p$. The requirement of covariance matrix stabilization, $\Sigma_1 = \ldots = \Sigma_q = \Sigma$ (Σ unspecified), as well as normalization seems realistic in light of the usual multivariate procedures. Following Dunn and Tubbs (1980), the joint likelihood for the data set $\{\underline{y}_{ij}\}$ can be re-expressed as the concentrated likelihood function

$$L(\lambda_1, \ldots, \lambda_p) = \exp(-np/2)(n/2\pi)^{np/2} \Phi(\lambda_1, \ldots, \lambda_p)^{-n/2}, \tag{5}$$

where

$$\Phi(\lambda_1, \ldots, \lambda_p) = |G| / \left(\prod_{k=1}^{p} \lambda_k \dot{y}_k^{\lambda_k - 1}\right)^2, \tag{6}$$

$$G = \sum_{i=1}^{q} \sum_{j=1}^{n_i} (\underline{z}_{ij} - \underline{z}_i)(\underline{z}_{ij} - \underline{z}_i)'$$

$$\underline{z}_{ij} = (y_{ij1}^{\lambda_1}, \ldots, y_{ijp}^{\lambda_p})',$$

$$\underline{z}_i = n_i^{-1} \sum_{j=1}^{n_i} \underline{z}_{ij},$$

$$\dot{y}_k = \left(\prod_{i=1}^{q} \prod_{j=1}^{n_i} y_{ijk} \right)^{1/n} \qquad (k = 1, \ldots, p).$$

Dunn and Tubbs (1980) have successfully applied the Fletcher-Powell conjugate gradient algorithm to obtain an iterative minimization of (6) with respect to $\{\lambda_k\}$, equivalent to maximizing (5), using a FORTRAN IV program locally known as VARSTB. Elements of the required gradient vector

$\nabla\Phi(\lambda_1,\ldots,\lambda_p)$ are given in that paper.

Example. Fisher's classical iris data (1936) consists of 50 observations from each of three varieties of iris, virginica, versicolor, and setosa. Each experimental unit consists of measurements of sepal length and width and petal length and width. This data is often used to illustrate classification and discriminant functions, implicitly assuming equality of the covariance matrices. If a check is performed, Bartlett's test gives $\chi^2 = 144.0$ with 20 degrees of freedom ($P < 0.0001$). VARSTB, on an IBM 370/155, converged to estimates $\hat{\lambda}_1 = -0.43054$, $\hat{\lambda}_2 = 0.51697$, $\hat{\lambda}_3 = 0.39843$, and $\hat{\lambda} = 0.55468$ in three iterations, requiring 47.02 seconds, starting with all exponents identically one. At convergence, the maximum stress between successive iterations was 3.3×10^{-14}. Retesting the transformed values, Bartlett's statistic was reduced to $\chi^2 = 65.2$. Even though still $P < 0.0001$, the reduction of Bartlett's statistic implies that the transformations have been effective. The necessity for transformations is substantiated by testing H_0: $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$, using the generalized likelihood ratio test

$$\chi^2 = n \ln[\Phi(1,1,1,1)/\Phi(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3,\hat{\lambda}_4)]$$
$$= 150 \ln[1.5497 \times 10^4/1.30735 \times 10^4]$$
$$= 25.51 \text{ with } 4 \text{ degrees of freedom}$$

($P<0.00005$). The effectiveness of the transformation will be examined further in a later section.

In summary, this procedure seems to satisfy the three criteria very well. It is based on the multi-normal likelihood function whose covariance matrix is independent of the mean (homoscedastic in the q-sample case). We may observe the decrease of Φ to its minimum in successive iterations, each step automatically determined by the last, and test the effectiveness of the proposed transformations at the end. Finally, the identities of the original variables are maintained, since each is simply raised to some power followed by a shift of location and a change in scale. Jackknifing is a logical means of examining the stability of the estimated powers as well as improving their robustness. Available computer time may become a limiting factor, however, if extensive jackknifing is attempted.

TRANSFORMATIONS TO UNIFORM (U) STATISTICS

Because of their simplistic distributional properties, transformations to uniform (U) statistics have universal appeal. Gnanadesikan (1977) summarized a variety of graphical methods for assessing marginal and joint normality, based on approximate transformations to i.i.d. U statistics. The exactness of his procedures will always be suspect, however, in that no compensation is made for replacing true mean vectors and covariance matrices by sample estimates. O'Reilly and Quesenberry (1973) demonstrated that by conditioning on sufficient statistics, the conditional distributions of a properly transformed subset of the original variables are exactly i.i.d. uniform, U(0,1). Because their results show the most promise due to their exactness and because they are scattered through a relatively recent literature, they are detailed here so far as they apply to assessing both marginal and joint normality in single and q-sample problems. The actual transformations are considered first, followed by a consideration of omnibus means of assessing uniformity. Since it will be required so frequently in the remainder of this section, let $G_\upsilon(\cdot)$ denote the cumulative distribution function (c.d.f.) of student's t distribution with υ degrees of freedom.

One-sample univariate. Suppose that $y_1,\ldots,y_n$ are i.i.d. $N(\mu,\sigma^2)$, and let $\bar{y}_r = \sum_{j=1}^{r} y_j/r$ and $s_r^2 = \sum_{j=1}^{r} (y_j - \bar{y}_r)^2/r$ for $r = 3,\ldots,n$. Then the $n-2$ random variables

$$u_{r-2} = G_{r-2}\{(r-2)^{1/2}(y_r-\bar{y}_r)/[(r-1)s_r^2 - (y_r-\bar{y}_r)^2]^{1/2}\} \quad (7)$$

are i.i.d. U(0,1) for $r = 3,\ldots,n$ (O'Reilly and Quesenberry 1973). In applications, one must guard against any systematic arrangement of the data (e.g., ordered from small to large) before applying the transformation given by (7).

Q-Sample univariate (heteroscedastic). Suppose that a random sample $y_{i1},\ldots,y_{i,n_i}$ of size n_i is drawn independently from each of q $N(\mu_i,\sigma_i^2)$ populations ($i=1,\ldots,q$). Then a set of $\sum_{i=1}^{q} (n_i - 2)$ i.i.d. U(0,1) random variables will result by combining the results of applying equation (7) separately to each sample (Quesenberry et al. 1976). The homoscedastic case will be treated by means of an analogous result in terms of the general linear model.

Univariate regression. Suppose that $\underline{y}_n \sim N(X_n\underline{\beta},\sigma^2 I)$ specifies the usual general linear model of full rank under homoscedastic, normality assumptions, where $\underline{\beta}$ is a $p \times 1$ vector of regression coefficients to be estimated. Let $\underline{x}_r'$ denote the r'th row and X_r denote the first r rows

of the overall design matrix X_n, $r = p+2,\ldots,n$. Provided that X_r is full column rank, let $c_r = \underline{y}_r'[I - X_r(X_r'X_r)^{-1}X_r']\underline{y}_r$ denote the residual sum of squares after fitting the model $\underline{y}_r = X_r\underline{\beta} + \underline{e}_r$. Then the $n - p - 1$ random variables $u_{r-p-1} = G_{r-p-1}(v_r)$, where

$$v_r = \frac{(r-p-1)^{1/2}[y_r - \underline{x}_r'(X_r'X_r)^{-1}X_r'\underline{y}_r]}{\{[1-\underline{x}_r'(X_r'X_r)^{-1}\underline{x}_r]c_r - [y_r - \underline{x}_r'(X_r'X_r)^{-1}X_r'\underline{y}_r]^2\}^{1/2}} \qquad (8)$$

are i.i.d. U(0,1) for $r = p+2,\ldots,n$ (O'Reilly and Quesenberry 1973). In applications, one must avoid any systematic arrangement of the data while at the same time insuring that X_{p+1} is of full rank.

Q-sample univariate (homoscedastic). Suppose that a random sample $y_{i1},\ldots,y_{i,ni}$ of size n_i is drawn independently from each of q, homoscedastic $N(\mu_i,\sigma^2)$ populations $(i = 1,\ldots,q)$. Let $n = \sum_{i=1}^{q} n_i$ represent the total sample size. Noting that the linear model $y_{ij} = \mu_i + e_{ij}$ for 1-way ANOVA is full rank, then the previous result applies provided that X_{q+1} contains at least one randomly selected observation from each of the q populations. If $\{y_{r(ij)}\}$ represents a randomized order across all samples of the remaining n-q-1 observations, let $\bar{y}_{(i)}$ represent the sample mean based on $n_{(i)}$ observations from the i'th sample at the r'th step of transformation (8) $(i = 1,\ldots,q)$. Let c_r be the corresponding error sum of squares from 1-way ANOVA at the r'th step. Then the n-q-1 random variables

$$u_{r-q-1}^{(ij)} = G_{r-q-1}\{(r-q-1)^{1/2}(y_{r(ij)} - \bar{y}_{(i)})/$$

$$[(1-n_{(i)}^{-1})c_r - (y_{r(ij)} - \bar{y}_{(i)})^2]^{1/2}\} \qquad (9)$$

are i.i.d. U(0,1) for $r = q+2,\ldots,n$. It will usually be informative to reassociate the random variables in their original sample groups once the above transformation has been made. In applications, the most common failure will be to forget to mix all the samples together in some random order before applying transformation (9).

One-sample multivariate. Suppose that $\underline{y}_1,\ldots,\underline{y}_n$ are i.i.d. from a full rank, p-variate $N(\underline{\mu}\ \Sigma)$ population. Let

$$\bar{\underline{y}}_r = r^{-1}\sum_{i=1}^{r}\underline{y}_i, \quad C_r = \sum_{i=1}^{r}\underline{y}_r\underline{y}_r' - r\bar{\underline{y}}_r\bar{\underline{y}}_r',$$

$$A_r'A_r = C_r^{-1}$$

(a Cholesky decomposition),

$$\underline{z}_r = A_r(\underline{y}_r - \bar{\underline{y}}_r)/[1 - 1/r - (\underline{y}_r - \bar{\underline{y}}_r)'C_r^{-1}(\underline{y}_r - \bar{\underline{y}}_r)]^{1/2}$$

$$= (z_{r1}, z_{r2}, \ldots, z_{rp})' \quad (r = p+2,\ldots,n).$$

Then the p(n-p-1) random variables given by

$$u_{r,s} = G_{r-p+s-2}\{z_{rs}[(r-p+s-2)/(1 + z_{r1}^2 + \ldots + z_{r,s-1}^2)]^{1/2}\} \qquad (10)$$

for $r = p+2,\ldots,n$ and $s = 1,\ldots,p$ are i.i.d. U(0,1) (Rincon-Gallardo et al. 1979).

Clearly, the multivariate, q-sample heteroscedastic case can be treated analogously to the univariate case using (10).

Q-sample multivariate (homoscedastic). Suppose that a random sample $\underline{y}_{i1},\ldots,\underline{y}_{i,n_i}$ of size n_i is drawn independently from each of q, homoscedastic, p-variate $N(\underline{\mu}_i,\Sigma)$ populations $(i = 1,\ldots,q)$. Let $n = \sum_{i=1}^{q} n_i$. Proceeding as in the q-sample univariate case, let us select p + q of the observations at random in such a way that at least one observation is selected from each of the q samples, and obtain a randomized order $\{\underline{y}_{r(ij)}\}$ across the samples of the remaining n-p-q observations. Let $\bar{\underline{y}}_{(i)}$ represent the sample mean vector based on $n_{(i)}$ observations at the r'th step of the transformation, and let E_r be the error matrix of SS and SP obtained from 1-way MANOVA at the r'th step. If we obtain the Cholesky decomposition $A_r'A_r = E_r^{-1}$ and define

$$\underline{z}_r = A_r(\underline{y}_{r(ij)} - \bar{\underline{y}}_{(i)})/$$

$$[1-n_{(i)}^{-1} - (\underline{y}_{r(ij)} - \bar{\underline{y}}_{(i)})'E_r^{-1}(\underline{y}_{r(ij)} - \bar{\underline{y}}_{(i)})]^{1/2},$$

$$= (z_{r1}, z_{r2}, \ldots, z_{rp})' \ ,$$

then by an extension of results given by Rincon-Gallardo et al. (1979), it follows that the p(n-p-q) random variables

$$u_{r,s}^{(ij)} = G_{r-p-q+s-1}\{z_{rs}[(r-p-q+s-1)/(1 + z_{r1}^2 + \ldots + z_{r,s-1}^2)]^{1/2}\} \qquad (11)$$

for $r = p+q+1,\ldots,n$ and $s = 1,\ldots p$ are i.i.d. U(0,1). Again, it will usually be informative to reassociate these random variables in their original sample groups once the transformations are completed.

Assessment of Uniformity

Suppose that $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(m)}$ represents the ordered values of m uniform statistics produced by any of the foregoing transformations. Under normality assumptions, the expected value of $u_{(j)}$ is $j/(m+1)$ for $j = 1,\ldots,m$, so that if we plot $u_{(j)}$ vs. $j/(m+1)$ in Cartesian coordinates, then we should expect the points to fall closely about the straight line defined by $g(u) = u$. In q-sample problems, it will generally be informative to assign different symbols to the points associated with the separate samples, or produce separate plots, in making a visual evaluation.

More formally, any of the standard, goodness-of-fit procedures (Pearson's χ^2, Kolmogorov-Smirov one-sample, etc.) can be applied. Miller and Quesenberry (1975) have shown that a modified Watson's U^2 statistic proposed by Stephens (1970)

$$U^2 = (12\ m)^{-1} + \sum_{j=1}^{m} [u_{(j)} - (j - 1/2)/m]^2 - m(\bar{u} - 1/2)^2 , \qquad (12)$$

where $\bar{u} = \sum_{j=1}^{m} u_{(j)}/m$, has attractive power properties as an omnibus test of uniformity. It has the additional advantage of having approximately constant 10, 5, 2.5, and 1 percentage points of 0.152, 0.187, 0.221, and 0.267 respectively for all $m \geq 10$ under the null assumption. The null hypothesis is rejected in this case for large values of U^2. This test in association with transformations given by equations (10) or (11) provides the only known exact test for multi-normality in either the one or q-sample cases. The only apparent disadvantage of these procedures is that the results are somewhat dependent on the order in which the observations are transformed, as we shall see in the following example.

Example. In order to study the effects of data presentation, 24 random permutations of Fisher's iris data, both with and without power transformations were subjected to the q-sample transformation defined by equation (11). The first $p + q = 4 + 3 = 7$ observations for each trial consisted of three randomly chosen versicolor and two each of virginica and setosa.

Table 2. Comparison of U^2 statistics for testing joint normality/homoscedasticity of Fisher's iris data with and without power transformations, using 24 random permutations of the data.

Randomization	Transformation No (U^2)	Transformation Yes (U^2)
1	0.470	0.248
2	0.110	0.084
3	0.243	0.115
4	0.126	0.031
5	0.354	0.130
6	0.364	0.128
7	0.083	0.054
8	0.147	0.105
9	0.603	0.217
10	0.270	0.117
11	0.057	0.060
12	0.260	0.119
13	0.254	0.095
14	0.118	0.115
15	0.199	0.047
16	0.229	0.060
17	0.215	0.059
18	0.218	0.058
19	0.163	0.030
20	0.260	0.174
21	0.354	0.185
22	0.072	0.063
23	0.124	0.069
24	0.251	0.153
Mean	0.231	0.105
s.d.	0.130	0.058

$U^2_{0.05} = 0.187$ $U^{2.025}_0 = 0.221$ $U^2_{0.01} = 0.267$

Table 2 summarizes the U^2 statistics resulting from equation (12) for testing the null hypothesis of joint normality and homoscedasticity. Clearly, VARSTB has had a beneficial impact, as reflected by the reduction of U^2 in every case. Figures 1a -1c show plots of the order statistics from randomization #21 for each variety before transformations were made, while figures 1d - 1f give the analogous results following transformations. It is evident that the setosa values are most affected. Applying the one-sample result of equation (10) to the setosa data alone yielded $U^2 = 0.236$ $(0.01 < P < 0.025)$ before, and $U^2 = 0.208$ $(0.025 < P < 0.05)$ after the transformations. Doubtless U^2 could be reduced still further if VARSTB were applied to the setosa data in isolation from the versicolor and virginica data.

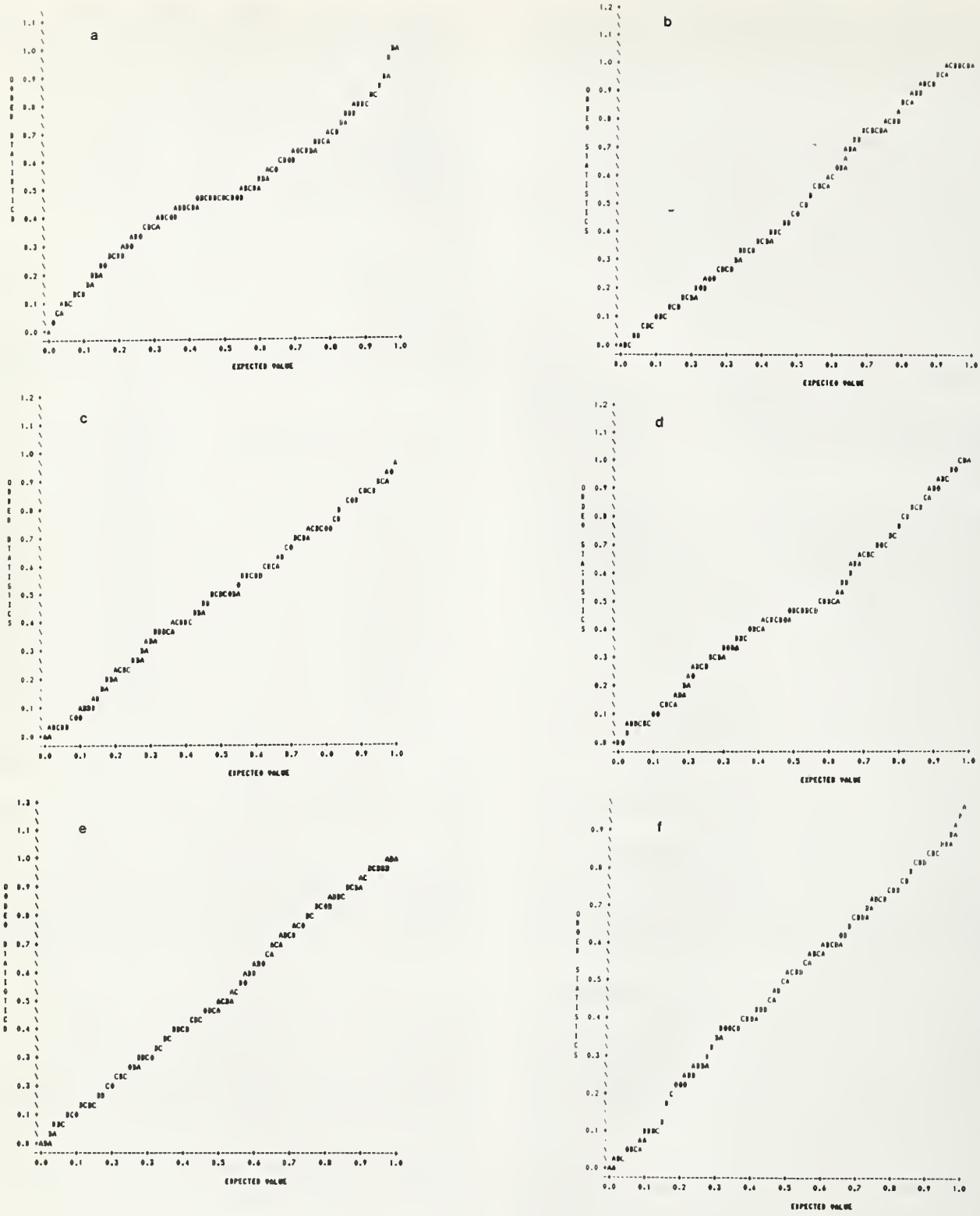


Figure 1. Plot of order statistics for Fisher's iris data: (a) original sertosa (b) original virginica (c) original versicular (d) power transformed sertosa (e) power transformed virginica (f) power transformed versicolor. A = 1 observation, B = 2 observations, C = 3 observations.

Table 3. Observed acceptance rates ($\alpha = 0.05$) for Bonferroni tests of joint normality/homoscedasticity on Fisher's iris data with and without power transformations, based on 24 random permutations of the data.

Summary	Transformation					
	No			Yes		
	Trials/test			Trials/test		
	1	2	5	1	2	5
No. of tests	24	23	20	24	23	20
No. of acceptances	9	5	6	22	22	19
Acceptance rate	0.375	0.217	0.300	0.917	0.957	0.950

A disconcerting finding in table 2 was the extreme variability of U^2 due simply to rearrangements in the order of the data. For the untransformed cases, U^2 was observed to range from a nonsignificant 0.057 to a highly significant 0.603. As summarized in table 3, the null hypothesis acceptance rates ($\alpha = 0.05$) were 0.375 and 0.917 respectively for the original and the transformed data. In order to produce a test of greater overall power in the presence of this lack of symmetry, the logical recommendation is to subject s random permutations of the data to the foregoing procedures. If a signficance level of α/s is used in each trial, and we reject the overall hypothesis if a rejection occurs for any trial, then the familiar Bonferroni result guarantees an overall significance level of at most α. Table 3 summarizes empirical acceptance rates for the Bonferroni procedure with $s = 2$ and $s = 5$ applied to the U^2 statistics of table 2, in the order shown. When $s = 2$, each successive pair was compared to $U^2_{0.05/2} = U^2_{0.025} = 0.221$ for an overall 0.05 level test. The analogous procedure for $s = 5$ employed $U^2_{0.05/5} = U^2_{0.01} = 0.267$. (In actual practice, either exactly $s = 2$ or $s = 5$ permutations would be tested.) Clearly, some increase in power is suggested by the decrease in acceptance rates, particularly when changing from $s = 1$ to $s = 2$ in terms of the untransformed data.

ACKNOWLEDGMENTS

The author wishes to thank Mr. Mike Meredith for sharing his SAS macro MULFIT for the computations involving Fisher's iris data. This work was partially supported by the Department of Energy Contract No. EY-76-S-05-5147.

LITERATURE CITED

Andrews, D.F., R. Gnanadesikan, and J.L. Warner. 1971. Transformations of multivariate data. Biometrics 27:825-840.

Andrews, D.F., R. Gnanadesikan, and J.L. Warner. 1973. Methods of assessing multivariate normality. p. 95-116 In Krishnaiah, P.R., editor. Multivariate Analysis III. Academic Press, New York, N.Y.

Anscombe, F.J. 1948. The transformation of Poisson, binomial, and negative binomial data. Biometrika 35:246-252.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society (B) 26:211-252.

Cockrell, D. 1980. SAS PROCEDURE RECODE. Computer Services Division, University of South Carolina. (Private communication.)

Cox, D.R., and N.J.H. Small. 1978. Testing multivariate normality. Biometrika 65:263-272.

Dunn, J.E., and G. Cappy. 1979a. A class of invertible functions for analysis of categorical response using linear models. p.182-187. In Fourth Annual SAS Users' Group International Conference. SAS Institute, Raleigh, N.C.

Dunn, J.E., and G. Cappy. 1979b. A class of invertible functions for analysis of categorical response using linear models. University of Arkansas Statistical Laboratory Technical Report No. 11.

Dunn, J.E., and J. Tubbs. 1980. VARSTB: A procedure for determining homoscedastic transformations of multivariate normal populations. Communications in Statistics Simulation and Computation B9(6):589-598.

Efron, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. Journal of American Statistical Association 70:892-898.

Gnanadesikan, R. 1977. Methods for statistical data analysis of multivariate observations. 311 p. John Wiley and Sons, New York, N.Y.

Grizzle, J.E., C.F. Starmer, and G.G. Koch. 1969. Analysis of categorical data by linear models. Biometrics 25:489-504.

Kendall, M.G., and A. Stuart. 1967. The advanced theory of statistics II. 437 p. Hafner Publishing Co., New York, N.Y.

Miller, F.L., and C. Quesenberry. 1975. Statistics for testing uniformity on the unit interval. Technical Report UNCCNC-CSD-12, Union Carbide Corporation, Nuclear Division.

Mosteller, F., and J. Tukey. 1977. Data analysis and regression. 588 p. Addison-Weley Publishing Company, Reading, Mass.

O'Reilly, F., and C.P. Quesenberry. 1973. The conditional probability integral transformation and applications to obtain composite chi-square goodness-of-fit tests. Annals of Statistics 1:74-83.

Press, S.J., and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. Journal of American Statistical Association 73:699-705.

Quesenberry, C.P., T.B. Whitaker, and J.W. Dickens. 1976. On testing normality using several samples: an analysis of peanut aflatoxin data. Biometrics 32:753-759.

Rao, C.R. 1952. Advanced statistical methods in biometric research. 390 p. John Wiley and Sons, New York, N.Y.

Rincon-Gallardo, S., C.P. Quesenberry, and F.J. O'Reilly. 1979. Conditional probability integral transformations and goodness-of-fit tests for multivariate normal distributions. Annals of Statistics 7:1052-1057.

SAS Institute. 1979. SAS users' guide. 494 p. Raleigh, N. Car.

Stephens, M.A. 1979. Use of the Kolmogorov-Smirov, Cramer-von Mises and related statistics without extensive tables. Journal of Royal Statistical Society (B) 32:115-122.

Applications: Ecological Theory, Habitat Management, Inventory

MULTIVARIATE ANALYSIS OF NICHE, HABITAT, AND ECOTOPE[1]

Andrew B. Carey[2]

Abstract.--Comprehension of the components of a species' response to an environmental complex can be achieved best by partitioning the full range of environmental factors potentially affecting the species (the ecotope hyperspace) into intercommunity factors (the habitat hyperspace) and intracommunity factors (the niche hyperspace). Hyperspaces, and the parts of hyperspaces occupied by a species (hypervolumes), are defined by multidimensional coordinates. Reduction in the number of dimensions of these hyperspaces and hypervolumes to increase comprehension of a species' response to them can be accomplished through multivariate analyses. The analysis of an ecosystem on a south-facing slope of the montane zone in Rocky Mountain National Park, Colorado is presented as an example. Principal component analysis was used to determine the major gradients in a habitat hyperspace defined by 21 environmental variables. Five principal components were interpreted as gradients of soil depth, soil moisture, ground cover, mammal distribution, and shrub abundance. Responses of five species of rodents to these gradients were determined by examining their relative positions on the principal components. Stepwise discriminant analysis (DA) was used to mathematically describe habitat and ecotope hypervolumes of the mammals. Comparison of the major determinants of each hypervolume of each species clarified the niche of the species. Furthermore, the distribution of a major ectoparasite of the mammals was analyzed by DA of environmental variables. Three levels of abundance of the wood tick Dermacentor andersoni could be predicted using only five environmental variables. The presence of a virus infecting both the ticks and mammals could be determined using seven environmental variables.

Key words: Discriminant analysis, ecotope, Eutamias, habitat, niche, principal component analysis, Spermophilus.

INTRODUCTION

Understanding the relationships between a species and its environment is the basic premise for wildlife management; to further this understanding is the foremost goal of wildlife research (Sanderson et al. 1979). Wildlife management and wildlife research have progressed to the point where precise and mathematical definition of basic terminology is necessary for further rapid progress. The response of a species (or ecotype) to its environment is usually examined in one of two contexts: the habitat (its response to extensive features) or the niche (its response to local features). Unfortunately, the communication of research results depends on the assumption of mutual understanding of terminology rather than on a basis of precise definition or an

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Research Wildlife Biologist, Northeastern Forest Experiment Station, USDA Forest Service, Morgantown, WV 26505.

organized system of mathematical descriptions. Haskell (1940) pointed out that this assumption is characteristic of a poorly advanced science.

Development of multidimensional treatments of habitat and niche (Hutchinson 1958) was concurrent with the adaptation of multivariate statistical methods to ecology (Hughes and Lindley 1955). These methods made practical the application of multidimensional thinking to field studies of species-environment relationships (Cody 1968, Green 1971, Hespenheide 1971, James 1971, Shugart and Patten 1972).

Despite these developments, there is still no unanimity on precise definitions of habitat and niche. Most ecology texts refer to habitat as the "address" and niche as the "profession" of an animal in its environment and later, in a discoursive context, refer to Hutchinson's n-dimensional hypervolume (Odum 1971, 1975; Kendeigh 1974; Smith 1974, 1977; Richardson 1977; Brewer 1979; McNaughton and Wolf 1979). Vandermeer (1972) pointed out that such definitions are excessively vague and inadequate. Whittaker et al. (1973) stated that the two terms, habitat and niche--perhaps the two most important in ecology--are among the most confused in usuage, and that their unsystematic usage has led to further confusion of other terms and concepts.

Nudds (1979) explained the need for applied wildlife research in a theoretical context, and Sanderson et al. (1979) reminded us that one of our major reasons for reporting wildlife research is to transfer that information from the research community to the management community. The need for clarity in interpretation and reports of research results and for some degree of unanimity on terminology is apparent. Whittaker et al. (1973, 1975) proposed a precise terminology (fig. 1) and found support in Hutchinson (1978). Despite objections to this terminology (Kulesza 1975, Rejmanek and Jenik 1975), I believe it offers the precise definition required to maintain a desirable level of clarity in the presentation of species-environment relationships. In this paper I will first discuss the terminology of Whittaker et al. (1973) with some additions from Hutchinson (1978), and then illustrate how it can be applied to a field study of a complex ecosystem. Finally, I will discuss the utility of the terminology in wildlife research and in management.

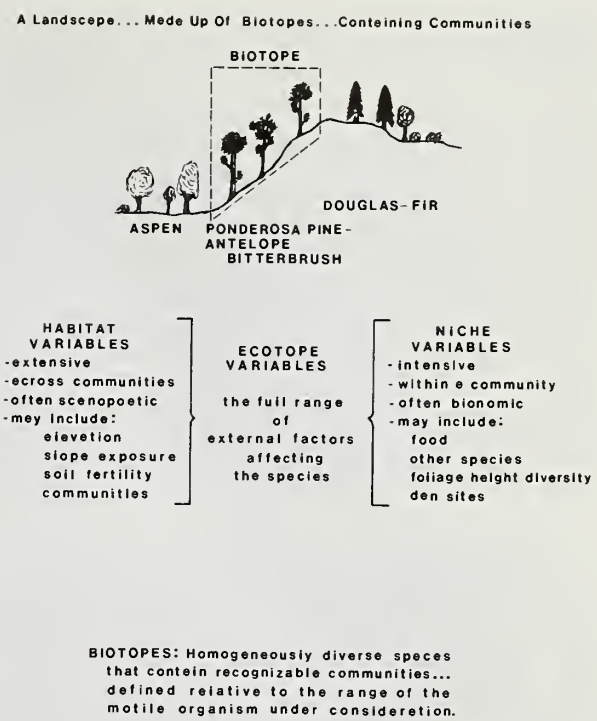


Figure 1. A suggested terminology.

NICHE, HABITAT, AND ECOTOPE

Landscapes are characterized by spatial gradients of structural (stage-setting or scenopoetic) variables (e.g., elevation, slope, soil fertility), and are composed of biotopes--locations (physical spaces) that have convenient arbitrary upper and lower boundaries, and that are horizontally homogeneously diverse (their structural elements are small compared with the range of an individual) relative to the larger motile organisms within them. These biotopes in the landscape pattern contain recognizable communities (fig. 1). Biological communities, however defined, are vaguely bound in space and time (Inger and Colwell 1977). The biotic context of a population is its community--an association of coexisting populations bound functionally by their interactions and spatially by their co-occurrence in a biotope (Colwell and Fuentes 1975).

The environmental variables with extensive spatial components are intercommunity or habitat variables; axes derived from these variables describe a multidimensional habitat hyperspace. Each species in the landscape occurs over some range of the habitat variables; the limits of these ranges define a habitat hypervolume--a fraction of the habitat hyperspace where a species occurs. Multivariate analyses can be used to reduce the number of dimensions defining hyperspaces and hypervolumes, and this allows comparison of the habitat relationships of the many species populations found in the habitat hyperspace. The response of a species population to habitat variables within its hypervolume describes its habitat; Maguire (1973) provided a method for examining such responses in terms of isopleths of population parameters projected on axes representing the reduced dimensions.

The intracommunity or niche variables (intensive or local environmental variables, e.g., height above ground, seasonal time, prey size, "microhabitat" variables) likewise define a niche

hyperspace that interrelates the species of a community. The niche hyperspace is analagous to the fundamental or preinteractive Hutchinsonian niche. Each species in the community differentially utilizes, occurs in, or is affected by some range of these variables; limits of these ranges define the niche hypervolume—a fraction of the niche hyperspace where the species occurs. The niche hypervolume is analagous to the realized or postinteractive Hutchinsonian niche including the "included niche" of Miller (1964). Niche variables are generally bionomic variables (resources for which there may be competition) and may define axes representing other member species of the community. The dimensions of niche hyperspaces and hypervolumes can be reduced through multivariate analyses, and isopleths of the population response to the niche variables (the niche) can be projected on them (Maguire 1973).

The landscape of communities (the compound hyperspace that represents the full range of external circumstances to which species in the landscape are adapted) is the ecotope hyperspace and contains an ecotope hypervolume defined by a species' limits on the ranges of the ecotope variables. The ecotope represents the full range of a species' adaptation to external factors and is the ultimate arena for consideration of its relations to its environment (especially in a broad evolutionary context), whereas the niche focuses on the role of the species within its community (especially competitive interactions), and the habitat relates to the distributional response of the species to the intercommunity environmental factors (those with extensive spatial components). Niche differences are intracommunity differences and involve genetic characteristics evolved in relation to other species; habitat differences reflect the evolutionary response to a gradient of environmental factors external to, although often modified by, the community. Niche breadth measures intrapopulation genetic characteristics; habitat breadth measures interpopulation differentiation based on mechanisms different from those responsible for niche breadth, for example, ecotypic and subspecific differentiation.

It can be seen readily that most wildlife management efforts do deal with habitat variables; however, the strength of the concepts is that they provide a conceptual scheme for different kinds of diversity (Hutchinson 1978) and, with increased interest in managing diversity (Pimlott 1969, Siderits and Radtke 1977), the wildlife biologist may want to deal with niche differences (within habitat or alpha-diversity), habitat differences (between habitats or beta-diversity), or with broad geographical differences (gamma-diversity).

The basic principle underlying multivariate analyses of species-environment relationships is the determination of points (limits of ranges of niche, habitat, or ecotope variables) in multidimensional space followed by the mathematical reduction of the number of dimensions so that those remaining are all orthogonal, independent, and significant (Hutchinson 1978).

AN ILLUSTRATION: A MONTANE ECOSYSTEM

The landscape chosen for study (Carey et al. 1980) included two south-facing slopes of the upper montane forest climax region in the Rocky Mountain National Park, near Estes Park, Colorado. The areas encompassed about 20 ha of varied terrain, ranging from steep (>30°) slopes with massive rock outcrops to gentle (2-8°) slopes with deep soil. Elevation was between 2,487 and 2,585 m, and the areas were representative of the upper montane, and contained all major plant communities found there: aspen stand complex, ponderosa pine complex, big sagebrush complex, and dry montane grassland complex (Marr 1967). Five species of rodents were abundant on the areas; in order of decreasing abundance they were deer mice (Peromyscus maniculatus), Richardson's ground squirrels (Spermophilus richardsonii), golden-mantled squirrels (S. lateralis), least chipmunks (Eutamias minimus), and Uinta chipmunks (E. umbrinus).

Methods

Sampling grids were surveyed on the two areas and consisted of 269 intersections (marked by stakes) 30 m apart (13x13 stakes on one area, 10x10 stakes on the other). A number of structural variables (table 1) were measured on the 225 30m x 30m quadrats covering the 20 ha. Seven plant-frequency counts were taken from 20cm x 20cm quadrats randomly placed in the vicinity of each stake. Four traps (for small mammals) were placed in the vicinity of each stake; these were operated for 46,000 trap nights. Fecal samples were collected from droppings beneath traps containing a chipmunk or ground squirrel. Fecal samples were analyzed by the microhistological method (Sparks and Malechek 1968) to obtain the dietaries of the species populations. Blood samples for Colorado tick fever virus isolation attempts were collected from a subsample of the captured rodents. Traps for adult, free-ranging ticks were placed in the center of 128 of the 225 quadrats and were operated for 1,200 trap nights. For detailed procedures see Carey et al. (1980).

Statistical Analyses

Structural variables that were highly correlated (>0.90) with simpler (not transgenerated or more easily measured) variables or that were invariant (e.g., DIRNSL) were deleted from the data set. The remaining variables were averaged over the sets of four contiguous quadrats surrounding each point location (stake) in the landscape. These variables were taken to represent habitat variables, and mean values for each point location were paired with values of the rodent-capture variables for that point location (total captures of the five most abundant

Table 1. Structural variables measured on 30m x 30m quadrats or transgenerated from measured variables.

Acronym	Structural variable
SOILD:	mean depth of soil exclusive of areas with < 5 cm of soil
%SOIL2:	percent of quadrat with soil < 5 cm in depth
%EXPRK:	percent of quadrat covered by exposed rock
RKRK:	rank (0-4)[1] of exposed rock for interstices
DSLOPE:	degree of slope
DIRNSL:	aspect (compass degrees)
GRASSL:	ranked abundance (0-4) of grass litter
PINEL:	ranked abundance (0-4) of pine needles and cones
WDDEBR:	ranked abundance (0-4) of wood debris less than 15 cm in diameter
LNFTLG:	length of logs > 15 cm diameter multiplied by diameter
DECOMP:	rank (0-4) of decomposition of log litter
SMPINE:	number of coniferous trees < 15 cm dbh
LGPINE:	number of coniferous trees > 15 cm dbh
NASPEN:	number of quaking aspen
%SHRUB:	percent of quadrat covered by shrubs
PUTR-P:	presence-absence (1,0) of antelope bitterbrush
RICE-P:	presence-absence (1,0) of wax currant
ARTR-P:	presence-absence (1,0) of basin big sagebrush
NJUSC:	number of Rocky Mountain junipers
DJUCO:	diameter (cm) of common junipers
ARLU-RK:	ranked abundance (0-4) of Louisiana sagewort
ARFR-RK:	ranked abundance (0-4) of fringed sagewort
ROAC-RK:	ranked abundance (0-4) of prickly rose
OPPO-RK:	ranked abundance (0-4) of plains pricklypear
PESI-RK:	ranked abundance (0-4) of mountain ball cactus
CHVI-P:	presence-absence (1,0) of Douglas rabbitbrush
EXPRKRK:	%EXPRK x RKRK
SHPUTR:	%SHRUB x PUTR-P
SHRICE:	%SHRUB x RICE-P
SHARTR:	%SHRUB x ARTR-P
SLPEXR:	DSLOPE x %EXPRK
TOPINE:	SMPINE + LGPINE
LNFTDC:	LNFTLG x DECOMP

[1]Rank (0-4): Absent, 0; sparse, 1; scattered, 2; common, 3; abundant, 4.

species). Principal component analysis (PCA) (Dixon 1976) was used to generate a smaller number of new variables (to reduce dimensions of the data set) that were interpreted as spatial gradients or major axes of the habitat hyperspace. Mammal-capture variables were included in the analysis to obtain a measure of the habitat, or response, of each species to the hyperspace gradients.

Point locations were grouped into species' ecotope and habitat hypervolumes and remaining hyperspaces according to the presence of each rodent species except for Richardson's ground squirrels and deer mice. The criterion for grouping stations in terms of Richardson's ground squirrels was that two or more ground squirrels had been captured near the point location. This criterion was used because single captures were scattered, and multiple-capture locations were clumped in distribution. The scattered captures presumably resulted from breeding season movements and dispersal behavior (Hansen 1962, Yeaton 1972, Michener and Michener 1977) and therefore, did not reflect habitat. Deer mice were trapped at virtually all locations.

The habitat hypervolumes of chipmunks and ground squirrels were mathematically described by stepwise discriminant analysis (DA) (Dixon 1976) of 21 of the habitat (scenopoetic) variables. The ecotope hypervolumes of the sciurids were similarly described by stepwise discriminant analysis of 9 habitat variables, 11 plant-frequency variables representing the distribution of the major food items of sciurids, and 4 mammal-capture variables (the four abundant species of rodents other than the species under consideration); these 24 variables collectively constitute ecotope variables. Results of the stepwise procedures (the best discriminating variables are chosen first) for the habitat and ecotope hypervolumes of a single species were compared to determine if the species was responding to habitat (scenopoetic) variables or to niche (bionomic) variables.

Distributions of adult ticks and virus in the landscape were described mathematically by discriminant analysis of structural variables (Carey 1979). The purpose of this analysis was to determine if enough information was contained in the structural variables to allow DA to generate discriminant functions (DF) that would be useful in classifying other parts of the landscape into categories of relative abundance of ticks and virus. A jackknife procedure (Brown 1977) was used to obtain less biased estimates of the error of classification of these discriminant functions.

Results

Principal Component Analysis

Five principal components (PC) were interpreted (table 2). These accounted for 64% of total variance in the data.

PC1 is a soil depth gradient in the habitat hyperspace. Measures of shallow soil, exposed rock, and slope were heavily weighted at one end with deep soil measures at the other end. CHVI-P,

ROAC-RK, and SRICH were at the deep soil end. Douglas rabbitbrush (Chrysothamnus viscidiflorus) was most abundant in areas of deep dry soil, and prickly rose (Rosa acicularis) was most abundant in areas of deep moist soil. Richardson's ground squirrels were abundant in areas of deep soil.

PC2 is a soil moisture gradient. Quaking aspen (Populus tremuloides), lush grass, and prickly rose were found on moist soil. Big sagebrush (Artemisia tridentata) grew only on deep dry soil, as did Douglas rabbitbrush. Fringed sagewort (Artemisia frigida) and shrub cover (antelope bitterbrush, Purshia tridentata, etc.) were well distributed over areas with dry soil.

PC3 is ground cover and abundant vegetation measures oppose litter measures. The order is vertical vegetative diversity, ranging from trees to shrubs to litter. PINEL was not highly correlated with TOPINE (r=0.16); pine litter accumulated in relatively closed stands and around scattered over-mature trees.

PC4 is mammal distribution in the habitat hyperspace. Richardson's ground squirrels (SRICH)are at one extreme of the component and the golden-mantled squirrels (SLAT) and chipmunks (TOTCH) aré at the other extreme; the deer mice (PMAN) are in the middle.

PC5 is a shrub abundance variable. Shrub measures are at one end and exposed rock measures are at the other end. The presence of CHVI-P seems to contradict the interpretation of PC5. However, CHVI-P was a presence-absence variable, not a measure of abundance. Douglas rabbitbrush was found in the absence of other shrubs in some deep dry soils and the stands of big sagebrush and antelope bitterbrush in other areas; it was rarely found in stands where antelope bitterbrush grew together with wax current (Ribes cereum) and boulder raspberry (Rubus deliciosus). TOPINE is associated with low shrub abundance in PC5; shrub cover was very low in pine stands.

Table 2. The principal components and their variable coefficients (from Carey et. al. 1980).

PC1--Soil depth		PC2--Soil moisture		PC3--Ground cover		PC4--Mammal distribution		PC5--Shrub abundance	
Variable	Coeff.	Variable	Coeff.	Variable	Coeff.	Variable	Coeff.	Variable	Coeff.
%SOIL2	0.36	NASPEN	-0.35	PINEL	-0.36	SLAT	0.37	SHARTR	-0.39
%EXPRK	0.31	ROAC-RK	-0.29	WDDEBR	-0.27	TOTCH	0.34	%SHRUB	-0.37
TOPINE	0.28	GRASSL	-0.24	LNFTLG	-0.23	WDDEBR	0.29	ARLU-RK	-0.36
EXPRKRK	0.26	NJUSC	-0.23	SLAT	-0.13	EXPRKRK	0.23	DJUCO	-0.31
ARLU-RK	0.26	SOILD	-0.22	CHVI-P	-0.07	%SHRUB	0.23	LNFTLG	-0.29
DSLOPE	0.24	DJUCO	-0.17	%SOIL2	-0.07	SHARTR	0.21	NJUSC	-0.21
NJUSC	0.23	PINEL	-0.15	DJUCO	-0.01	SOILD	0.17	ROAC-RK	-0.14
PMAN	0.17	TOTCH	-0.08	ARLU-RK	0.03	%EXPRK	0.17	WDDEBR	-0.15
LNFTLG	0.15	SRICH	-0.06	NJUSC	0.04	DJUCO	0.17	NASPEN	-0.09
SLAT	0.15	LNFTLG	-0.05	SOILD	0.08	CHVI-P	0.15	ARFR-RK	-0.09
WDDEBR	0.11	WDDEBR	-0.00	ARFR-RK	0.11	NASPEN	0.12	%SOIL2	-0.06
TOTCH	0.07	%SOIL2	0.00	%SHRUB	0.15	ROAC-RK	0.11	GRASSL	-0.05
ARFR-RK	0.04	%EXPPRK	0.01	SHARTR	0.16	PINEL	0.10	DSLOPE	0.00
PINEL	0.02	ARLU-RK	0.01	%EXPRK	0.19	PMAN	0.06	TOTCH	0.02
DJUCO	-0.05	PMAN	0.03	SRICH	0.18	LNFTLG	0.02	SRICH	0.05
SHARTR	-0.09	SLAT	0.05	TOTCH	0.19	GRASSL	0.02	SOILD	0.07
NASPEN	-0.10	EXPRKRK	0.07	EXPRKRK	0.20	TOPINE	0.02	PMAN	0.09
%SHRUB	-0.12	TOPINE	0.09	PMAN	0.22	NJUSC	0.01	PINEL	0.13
GRASSL	-0.14	DSLOPE	0.15	GRASSL	0.27	%SOIL2	-0.03	TOPINE	0.15
ROAC-RK	-0.17	CHVI-P	0.26	NASPEN	0.28	DSLOPE	-0.10	%EXPRK	0.19
SRICH	-0.22	ARFR-RK	0.33	DSLOPE	0.30	ARLU-RK	-0.16	SLAT	0.20
CHVI-P	-0.28	SHARTR	0.34	TOPINE	0.31	ARFR-RK	-0.35	CHVI-P	0.21
SOILD	-0.28	%SHRUB	0.34	ROAC-RK	0.32	SRICH	-0.35	EXPRKRK	0.27
Eigenvalue	6.3		3.6		2.2		2.1		1.7
Cumulative percent[1]	0.25		0.40		0.49		0.57		0.64

[1]Cumulative proportion of total variance explained by the principal components.

Thus, the major gradients in the habitat hyperspace were soil depth and soil moisture, and they accounted for two-thirds of the explained variance. Smaller portions of the variance were explained by ground cover, mammal distribution, and shrub abundance. Barkham and Norris (1970) pointed out that it is not uncommon for minor components to increase in complexity and to represent interaction effects. The last three components probably do, in part, represent the interaction of soil depth and soil moisture. PCA illustrated the plants' and mammals' responses (habitats) to the habitat hyperspace. Richardson's ground squirrels were strongly associated with deep soils (PC1). Golden-mantled ground squirrels were moderately associated with shallow soils (PC1) and exposed rock (PC1, PC5). Chipmunks were associated with golden-mantled squirrels (PC4). Deer mice were associated with grasses (PC3) and were intermediate on the other PC's. All of the mammals occupied intermediate positions on the soil moisture gradient. This indicates that soil moisture was not a major determinant of any species' habitat hypervolume and suggests that mammal distributions were a function of soil depth. Richardson's ground squirrels almost exclusively used ground burrows for escape and denning. They assumed a characteristic "picket pin" posture for observation. They did not use rocks for escape or observation in the study areas. Golden-mantled ground squirrels used large rocks as observation, feeding, and basking posts. They commonly used rock interstices for escape cover and den sites. Chipmunks also used rocks as observation posts, escape cover, and den sites.

Discriminant Analyses on Mammals

Results of DA on data sets of the habitat and ecotope hyperspaces are illustrated in table 3. The first five steps and the last step of the DA are shown. Group means of the habitat hypervolume and the remaining hyperspace were significantly different ($P<0.001$) at each step for all species except the least chipmunk. For the least chipmunk, $P=0.012$ at the first step and $P=0.019$ at the last step. Group means of the ecotope hypervolume and the remaining hyperspace were significantly different ($P<0.001$) at each step for all species. Rates of change in the U statistics and the percentages of stations correctly classified showed that most information contained in the DF's was contributed by the first few variables in each case. More stations were classified correctly with the ecotope hyperspace variables than with the habitat hyperspace variables. U statistics were lower and station classifications were more successful for golden-mantled ground squirrels and Richardson's ground squirrels than for Uinta chipmunks and least chipmunks.

Comparison of discriminating variables between the habitat hypervolumes and the ecotope hypervolumes showed that the ecotope hypervolume of the Richardson's ground squirrel differed little from its habitat hypervolume, and that its distribution was a response primarily to a structural variable, soil depth, and incidentally to a bionomic variable, captures of golden-mantled ground squirrels. The golden-mantled ground squirrel's distribution was a function of bionomic variables--captures of other mammals (primarily Richardson's ground squirrels)--rather than a response to habitat (structural) variables. Habitat variables (e.g., EXPRKRK) were equal in importance to bionomic variables (e.g., captures of golden-mantled ground squirrels) in describing the distribution of least chipmunks. The Uinta chipmunk responded to habitat variables.

Table 3. Results of stepwise discriminant analysis of hyperspaces for Spermophilus lateralis (modified from Carey et al. 1980).

Variable set	Step[1]		U	%
Habitat	1	%SOIL2	0.87	66
	2	WDDEBR	0.81	72
	3	CHVI	0.79	75
	4	PINEL	0.78	75
	5	EXPRKRK	0.77	76
	21	GRASSL	0.72	77
Ecotope	1	SRICH	0.67	86
	2	PMAN	0.65	86
	3	ARLU-F	0.64	87
	4	EMIN	0.63	87
	5	%SHRUB	0.62	86
	24	EXPRKRK	0.59	88

[1]Step in stepwise procedure, variables included, Wilks λ, cumulative percent of stations properly assigned. All group means signifincantly different at all steps ($P<0.01$).

Summary of Species-Environment Relationships

Richardson's Ground Squirrel.--Food habits of the Richardson's ground squirrels were similar to those of the golden-mantled ground squirrels. Spatial overlap between the two species in the habitat hyperspace was moderate, and ratios of the mean captures of the two in each other's habitat hypervolume were inversely related. Diet overlap with the least chipmunk was small; spatial overlap moderate. Richardson's ground squirrels were at one end of the mammal distribution component, and golden-mantled ground squirrels and chipmunks at

the other end. PCA and means of the discriminating variables showed that the habitat hypervolume of Richardson's ground squirrel was characterized by deep soil, vegetative ground cover, and slight slopes. It was not characterized by either extreme of soil moisture, or by high values of exposed rock or shrub cover. Discriminating variables for the ecotope hypervolume demonstrated the overwhelming influence of the habitat (structural) variables, especially soil depth measures, but also indicate a degree of negative response to the golden-mantled ground squirrels.

Golden-mantled Ground Squirrels.--Trophic overlap was great with Richardson's ground squirrel, and moderately low with the least chipmunk. Spatial overlap was small with Richardson's ground squirrel, and great with the least chipmunk. Capture ratios were inversely related with the ground squirrel, and one-sided with the chipmunk (in favor of the golden-mantled ground squirrel). Their habitat hypervolume was characterized by shallow, somewhat dry soil on steep slopes with sparse grass cover, moderate tree (ponderosa pine, Pinus ponderosa) cover, abundant litter, and abundant exposed rocks with interstices. The best discriminating variable for the ecotope hypervolume was the captures of Richardson's ground squirrel, a bionomic variable. Other discriminating variables contributed little to the description of the ecotope hypervolume.

Least Chipmunk.--Overlap between the diet of least chipmunks and those of ground squirrels was small; the least chipmunk made greater use of arthropods for food than did ground squirrels. Spatial overlap with the golden-mantled ground squirrel was high. The ground squirrel was near 92% of the point locations in the chipmunk habitat hypervolume and outnumbered chipmunks throughout the habitat hyperspace. The habitat hypervolume of the least chipmunk was characterized by intermediate soil depth and moisture, moderate vegetative ground cover, and higher than average values of exposed rock, rock interstices, prickly rose, aspen, pine trees (P. ponderosa and P. contorta), and common juniper (Juniperus communis). Exposed rock with numerous interstices was the single best descriptor of the habitat hypervolume, regardless of whether rocks were in conjunction with deep or shallow soils, pine trees or aspens, or shrubs or grass. The ecotope hypervolume was characterized by the presence of golden-mantled ground squirrels. Deep soil indicators (NASPEN, POPR-F, ARTR-F) were less important discriminating variables, and described the part of the hypervolume not characterized by golden-mantled ground squirrels, for example, small rock outcrops in deep soil areas.

Uinta Chipmunk.--Uinta chipmunks were relatively few in number and were caught in only 16% of the trap stations. Their habitat and ecotope hypervolumes were characterized by the highest mean values for shallow soil, exposed rock, rock interstices, steep slopes, log litter, and pine trees, and by the lowest mean values for soil depth, shrub cover, grass litter, and deep soil indicators, such as prickly rose and Douglas rabbitbrush.

Discriminant Analyses on Ticks and Virus

Ticks.--The relationship between habitat variables and adult wood tick abundance is illustrated in table 4. High tick abundance was associated with shallow soil, moderate shrub cover, abundant exposed rock and rock interstices, steep slopes, relatively abundant pine, and abundant log litter. Moderate tick abundance was associated with intermediate values of the same variables, except for shrub cover, which was abundant. The distribution of ticks in the habitat hyperspace reflects their physiological tolerance for soil temperatures and moisture (Wilkinson 1967).

The objectives of the DA of tick distribution were to obtain DF's of variables that were easily and quickly measured and that provided a high degree of discriminance in classifying areas on the basis of tick abundance. The discriminant functions in table 4 met those objectives. The two best discriminating variables were DSLOPE and %SOIL2. Less biased estimates of errors of misclassification were obtained with jackknifing procedures; they differed little from the cumulative percentages in table 4 (Carey 1979).

CTF Virus.--The relationship between virus activity and the ecotope hyperspace variables was illustrated by Carey (1979). Virus was isolated from mammals captured in areas characterized by shallow soil, abundant exposed rocks and rock interstices, relatively abundant pine, moderate shrub cover, moderately steep slopes, relatively high numbers of golden-mantled squirrels, least chipmunks, Uinta chipmunks, deer mice, and immature ticks, and low numbers of Richardson's ground squirrels. The objective of the DA of virus distribution was to obtain DF of easily measured variables that would permit identification of areas maintaining virus circulation with a low probability of misclassifying areas with virus. The DF met this objective (Carey 1979).

DISCUSSION

I believe the foregoing is a demonstration of precise definition combined with methods of multivariate analysis. After the terminology is grasped, the results should be ordered and clear to the reader. The distinction between structural (scenopoetic) variables and bionomic (potentially measuring competition) variables should be clear; also the differences in scope between habitat (intercommunity), niche (intracommunity), and ecotope variables should be clear. The enforced distinction between niche, habitat, and ecotope results in heuristic data analysis and interpretation of results that is clear. The compatibility between the terminology and the analysis is obvious.

Table 4. Discriminating variables and discriminant functions for the distribution of adult, free-ranging wood ticks. All group means significantly different at P<0.01 using approximate F tests.

Variable	U	Cumulative[1] %	DF Coefficients[2] None-Few ticks	DF Coefficients[2] Few-Many ticks
%SOIL2	0.519	62.5	0.21450	0.25643
DSLOPE	0.360	68.8	0.19518	0.44237
%SHRUB	0.324	72.7	0.24368	0.21132
ARLU-RK	0.302	72.7	1.16317	2.19884
SOILD	0.291	73.4	1.23637	1.21552
GRASSL	0.281	74.2	-0.17259	-0.78330
NASPEN	0.263	76.6	0.33857	2.63709
Constant			-14.55902	-21.57874

[1]Cumulative percentages of stations correctly assigned by the DF.
[2]Coefficients of the variables for the discriminant function separating the two groups.

These concepts (the ecological and the statistical) have been used to good advantage in the past, but, I believe, without precise definition. James (1971) clearly distinguished between intercommunity and intracommunity relationships in addition to formulating the niche-gestalt. Anderson and Shugart (1974) used the same (Whittaker et al. 1973) definition of habitat hyperspace. Dueser and Shugart (1978, 1979) also recognized the distinction between inter- and intracommunity relationships, but used a plethora of terms (habitat, microhabitat, habitat type, habitat association, habitat patch, patch type, forest stand type, forest type, desert community, rodent community, forest biome, fine-grained environment, etc.) in the 1978 paper (I am not criticizing the quality of their scientific contribution, I am pointing out the lack of precise terminology). Similarly M'Closkey (1975, M'Closkey and Fieldwick 1975) recognized the inter- and intracommunity distinction and the applicability of multivariate analysis (M'Closkey 1976) to the study of animal-environment relationships but not the use of precise terminology.

The use of the suggested terminology and multivariate analysis in the study of animal-environment relationships is clear, but what about the applicability to wildlife management? Aside from the basic understanding of how a species (or ecotope) responds to its environment, I think the inter- and intracommunity distinction is one that will prove to be of great value to the wildlife biologist, especially if the trend towards management for diversity continues. The distinction between structural and bionomic variables is useful at the management level; extensive management must focus on scenopoetic variables, but intensive management can consider bionomic variables. Understanding the concepts of spatial gradients and species' responses to them is fundamental to predicting results (or "environmental impact") of a management decision. The concept of homogeneously diverse biotopes coupled with inter- and intracommunity distinctions is of great value in determining scale or size for both sampling units and managing units.

The use of the terms "hyperspace" and "hypervolume" may be confusing and, therefore, of less benefit. Simpler terms, for example "preinteractive" (or fundamental) and "postinteractive" (or realized) could be used to modify "niche space" and "habitat space" without changing the concepts. Applying theoretical concepts to field studies can be rewarding. In the example given, discriminant functions for the ticks and virus could be used to locate recreational facilities and activities for minimum human exposure to the virus. The tick discriminant function could be used in the application of acaricides. Richardson's ground squirrel is a reservoir of plague (Yersinia pestis), and the golden-mantled ground squirrel is a reservoir of Colorado tick fever virus. Results of the analysis suggest that if Richardson's ground squirrels were controlled (killed) to minimize human exposure to plague (such a control effort was instituted in 1976 and 1977), they would be replaced by golden-mantled ground squirrels. This would not increase the risk of human exposure to Colorado tick virus, because the golden-mantled ground squirrels would have moved

out of the area that was climatically suited for ticks.

Many authors, especially Green, (1971, 1974) have discussed the applicability of multivariate analyses in ecology; most of the comments apply equally to wildlife research. However, I have seen only two reports in the Journal of Wildlife Management (Klebenow 1969, Martinka 1972) describing the application of multivariate analyses to animal-environment relationships. I believe a functional organization and a standardization of ecological terminology can do much to promote the application of multivariate statistics to the more applied areas of wildlife reasearch.

LITERATURE CITED

Anderson, S.H., and H.H. Shugart, Jr. 1974. Habitat selection of breeding birds in an east Tennessee deciduous forest. Ecology 55:828-837.

Barkham, J.P., and J.M. Norris. 1970. Multivariate procedures in an investigation of vegetation and soil relations of two beech woodlands, Cotswald Hills, England. Ecology 51:630-639.

Brewer, R. 1979. Principles of ecology. 299 p. W.B. Saunders Co., Philadelphia, Penn.

Brown, M.B., editor. 1977. BMDP-77. Biomedical computer programs. P-series. 880 p. University of California Press, Berkeley.

Carey, A.B. 1979. Discriminant analysis: a method of identifying foci of vector-borne diseases. American Journal Tropical Medicine and Hygiene 28:750-755.

Carey, A.B., R.G. McLean, and G.O. Maupin. 1980. The structure of a Colorado tick fever ecosystem. Ecological Monograph 50:131-151.

Cody, M.L. 1968. On the methods of resource division in grassland bird communities. American Naturalist 102:107-147.

Colwell, R.K., and E.R. Fuentes. 1975. Experimental studies of the niche. Annual Review of Ecology and Systematics 6:281-310.

Dixon, W.J., editor. 1976. B.M.D. Biomedical computer programs. 773 p. University of California Press, Berkeley.

Dueser, R.D., and H.H. Shugart, Jr. 1978. Microhabitats in a forest-floor small-mammal fauna. Ecology 59:89-98.

Dueser, R.D., and H.H. Shugart, Jr. 1979. Niche pattern in a forest-floor small-mammal fauna. Ecology 60:108-118.

Green, R.H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of central Canada. Ecology 52:543-556.

Green, R.H. 1974. Multivariate niche analysis with temporally varying environment factors. Ecology 55:73-83.

Hansen, R.M. 1962. Dispersal of Richardson's ground squirrel in Colorado. American Midland Naturalist 68:58-66.

Haskell, E.J. 1940. Mathematical systematization of "environment," "organism" and "habitat." Ecology 21:1-16.

Hespenheide, H.A. 1971. Flycatcher habitat selection in the eastern deciduous forest. Auk 88:61-74.

Hughes, R.E., and D.V. Lindley. 1955. Application of biometric methods to problems of classification in ecology. Nature 175:806-807.

Hutchinson, G.E. 1958. Concluding remarks. Cold Spring Harbor Symposium on Quantitative Biology 22:415-427.

Hutchinson, G.E. 1978. An introduction to population ecology. 260 p. Yale University Press, New Haven, Conn.

Inger, R.F., and R.K. Colwell. 1977. Organization of contiguous communities of amphibians and reptiles in Thailand. Ecological Monographs 47:229-253.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Kendeigh, S.C. 1974. Ecology with special reference to animals and man. 474 p. Prentice-Hall, Inc., Englewood Cliffs, N.J.

Klebenow, D.A. 1969. Sage grouse nesting and brood habitat in Idaho. Journal of Wildlife Management 33:649-667.

Kulesza. G. 1975. Comment on "Niche, habitat, and ecotope." American Naturalist 109: 476-479.

M'Closkey, R.T. 1975. Habitat succession and rodent distribution. Journal of Mammalogy 56:950-955.

M'Closkey, R.T. 1976. Community structure in sympatric rodents. Ecology 57:728-739.

M'Closkey, R.T., and B. Fieldwick. 1975. Ecological separation of sympatraic rodents (Peromyscus and Microtus). Journal of Mammalogy 56:119-129.

McNaughton, S.J., and L.L. Wolf. 1979. General ecology. 702 p. Hold, Rinehart and Winston, New York, N.Y.

Maguire, B., Jr. 1973. Niche response structure and the anlaytical potentials of its relationship to the habitat. American Naturalist 107:213-246.

Marr, J.W. 1967. Ecosystems of the east slope of the front range in Colorado. 134 p. University of Colorado Studies Series in Biology No. 8, Boulder, Colo.

Martinka, R.R. 1972. Structural characteristics of blue grouse territories in southwestern Montana. Journal of Wildlife Management 36:498-510.

Michener, G.R., and D.R. Michener. 1977. Population structure and dispersal in Richardson's ground squirrels. Ecology 58:359-368.

Miller, R.S. 1964. Ecology and distribution of pocket gophers (Geomyidae) in Colorado. Ecology 45:256-272.

Nudds, T.D. 1979. Theory in wildlife conservation and management. Transactions North American Wildlife and Natural Resources Conference 44:277-288.

Odum, E.P. 1971. Fundamentals of ecology. third edition. 574 p. W.B. Saunders Co., Philadelphia, Penn.

Odum, E.P. 1975. Ecology: the link between the natural and social sciences. 244 p. Holt, Rinehart and Winston, New York, N.Y.

Pimlott, D.H. 1969. The value of diversity. Transactions North American Wildlife and Natural Resources Conference 34:265-280.

Rejmanek, M., and J. Jenik. 1975. Niche, habitat, and related ecological concepts. Acta Biotheoreologica 24:100-107.

Richardson, J.L. 1977. Dimensions of ecology. 412 p. The Williams and Wilkins Co., Baltimore, Md.

Sanderson, G.C., E.D. Ables, R.D. Sparrowe, J.R. Grieb, L.D. Harris, and A.N. Moen. 1979. Research needs in wildlife. Transactions North American Wildlife and Natural Resources Conference 44:166-175.

Shugart, H.H., Jr., and B.C. Patten. 1972. Niche quantifiation and the concept of niche pattern. p. 283-325. In B.C. Patten, editor. Systems analysis and simulation in ecology. Volume 2. Academic Press, New York, N.Y.

Siderits, K., and R.E. Radtke. 1977. Enhancing forest wildlife habitat through diversity. Transactions North American Wildlife and Natural Resources Conference 42:425-434.

Smith, R.L. 1974. Ecology and field biology. 850 p. Second edition. Harper and Row Publishers, New York, N.Y.

Smith, R.L. 1977. Elements of ecology and field biology. 497 p. Harper and Row Publishers, New York, N.Y.

Sparks, D.R., and J.C. Malechek. 1968. Estimating percentage dry weight in diets using a microscope technique. Journal of Range Management 21:261-265.

Vandermeer, J.H. 1972. Niche theory. Annual Review of Ecology and Systematics 3:107-132.

Whittaker, R.H., S.A. Levin, and R.B. Root. 1973. Niche, habitat, and ecotope. American Naturalist 107-321-338.

Whittaker, R.H., S.A. Levin, and R.B. Root. 1975. On the reasons for distinguishing "niche, habitat, and ecotope." American Naturalist 109:479-482.

Wilkinson, P.R. 1967. The distribution of Dermacentor ticks in Canada in relation to bioclimatic zones. Canadian Journal of Zoology 45:517-537.

Yeaton, R.I. 1972. Social behavior and social organization in Richardson's ground squirrel (Spermophilus richardsonii) in Saskatchewan. Journal of Mammalogy 53:139-147.

DISCUSSION

BOB CLARK: Were there age and sex biases in the captures?

A.B. CAREY: Possibly. I attempted to minimize biases due to age and sex by disregarding single captures of Richardson's ground squirrels, to trap-happiness by only using the first capture of an individual in a sampling period, and to competition for traps by placing four traps at each point location in the landscape and by checking the traps up to four times daily.

BOB CLARK: By removing biases due to dispersal, would you improve the discrimination of the habitats?

A.B. CAREY: Yes. However, the sample size was large and dispersal movements probably accounted for a small proportion of the captures used as captures seemed to be very clumped. In regards to other biases, it should be emphasized that I was interested only in generating a relatively crude measure of population response with the objective of determining hypervolumes.

KEN MORRISON: Given that there are behavioral biases inherent in trapping data, would this bias any results and/or interpretations of such an analysis?

A.B. CAREY: Yes, but see previous response to Bob Clark.

R. DUESER: How important is the notion of the "homogeneously diverse biotope" to your distinction between "niche" and "habitat"? What would be the implications of the non-existence of homogeneously diverse biotopes?

A.B. CAREY: The distinction between niche and habitat is that between intra- and inter-community, thus one must be able to assign a point location in the landscape to a recognizable community which occupies the homogeneously diverse space called the biotope, thus the notion is fundamental. "Homogeneously diverse" is relative to the range of a motile organism; if communities exists, so do biotopes; if they do not, niches don't either.

JAMES DUNN: (1) If the principal components are properly named, can variation in soil really be independent of moisture? (2) Can variation in mammal distribution really be independent of variation in ground cover? (3) Why not confess that the fundamental variables need not be independent and proceed with an oblique factor solution. The additional benefit might be a measure of the association between say mammal factor and cover factor if essentially the same named factors result?

A.B. CAREY: (1) Soil depth and soil moisture can be quite independent as is reflected by the community pattern in the landscape e.g., basin-big sagebrush on deep dry soils, and aspen stand complexes on deep wet soils. (2) Variation in mammal distribution can be independent of "ground cover" especially in a homogeneously diverse space. (3) Principal components beyond the 1st and 2nd components may show interaction effects (Barkham & Norris 1970) as mine do. However, I think my principal components nicely describe the spatial gradients in the landscape. I would say den sites (ground dens, rock dens) were more important than vegetative cover; there is also good evidence that food is not limiting and that the behavior of the species causes them to avoid the densest cover.

FORHAB: A FOREST SIMULATION MODEL TO PREDICT HABITAT STRUCTURE FOR NONGAME BIRD SPECIES[1]

T.M. Smith[2], H.H. Shugart[3], and D.C. West[4]

Abstract.--FORHAB (a deciduous forest stand stimulation model) was used to predict changes in available breeding habitat for the avian community inhabiting the Walker Branch Watershed in east Tennessee. A census was conducted to locate all breeding territories of the various bird species on the watershed. Data on vegetational structure of these territories were used to calculate linear decision scales, a classification procedure based on discriminant function analysis, which could be used to classify forest stands as potential breeding habitat for the various bird species. FORHAB was used to simulate changes in forest structure of the watershed due to both natural succession and certain introduced forest management practices (diameter-limit cut). Variables describing the vegetational structure of the forest stands generated by FORHAB were used to determine availability of potential breeding habitat for each bird species through time using a subroutine based on the above-mentioned classification procedure. Predictions of available habitat for the ovenbird (Seiurus aurocapillus) on the Walker Branch Watershed are presented as an example of model output.

Key words: Avian community; deciduous forest; discriminant function analysis; forest management; linear decision scales; nongame birds; simulation model; vegetation structure.

INTRODUCTION

The effects of modern forest management practices on wildlife has been a subject of growing concern in recent years (Slusher and Hinckley 1974, Thomas 1979). A major portion of this interest has been directed toward the nongame bird component of an animal community (Smith 1975, DeGraaf 1978) particularly the effect that management-related disturbances have on the population dynamics of either a single species (e.g., endangered species) or the avian community as a whole. The majority of work in this area has been of a descriptive nature. Researchers have dealt primarily with general habitat preferences of selected species (fig. 1) or the role of structural heterogenity on the overall avian diversity of the forest (MacArthur and MacArthur

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25,1980, Burlington Vt.

[2]Graduate Research Fellow, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge TN 37830.

[3]Senior Research Staff Member, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830.

[4]Research Staff Member, Envrionmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge TN 37830.

1961). This information, coupled with a knowledge of effects of various silvicultural practices on structure of the forest, has been the basis of timber-wildlife management decisions to date. Problems of time and expense have limited the number of actual field studies which have monitored directly the effects of timber management on bird populations (Hagar 1960, Franzreb and Ohmart 1978). However, even these studies are of limited value. Site-specific conditions and past history of the forest greatly limit the generalization of results from a specific study to other forests.

Recently, techniques of multivariate statistics have allowed analyses beyond qualitative methods of describing habitat preferences of avian species. With the use of multivariate statistical procedures, one can classify forest stands quantitatively with respect to habitat potential. Multivariate data analysis has been used to quantify the microhabitat selection patterns at both the species and community level (James 1971, Shugart and Patten 1972, Anderson and Shugart 1974, Whitmore 1975). Conner and Adkisson (1976) proposed a method of classifying forest stands as suitable woodpecker habitat using a discriminant function analysis procedure based on variables describing structure of forest vegetation and stressed the potential of such techniques as management tools. However, applicability of such statistical procedures is hindered by lack of habitat data necessary for dynamic habitat analyses. Management is by definition a dynamic process. To assess potential effects of various management strategies on availability of habitat, quantitative data describing changes in structure of the forest through time is necessary. To date, quantitative information on structural changes of the forest resulting from various timber management practices is not available. A forest stand simulation model may very well remedy problems inherent in assessing habitat modifications associated with various management techniques.

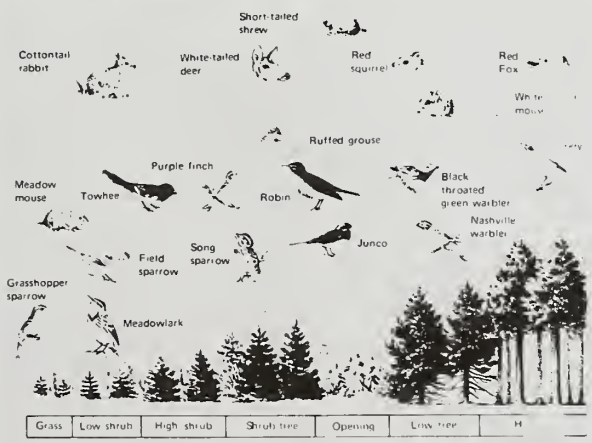


Figure 1. General habitat preference of various animal species inhabiting a northeastern conifer forest. [Reprinted with permission from Smith (1980)].

Forest stand simulators (Shugart and West 1980), can be used to assess effects of alternate forest management strategies on selected bird species or for entire avian communities. Models vary in their mathematical structure, but typically function by considering tree-by-tree changes over time for an area that corresponds to that of a canopy tree or some sample unit. The spatial scale of these models corresponds to what has been termed the microhabitat scale for birds. Because such simulators have the ability to predict structural changes in the forest through time, one can couple this quantitative data with statistical classification procedures previously mentioned to project long-term consequences of different management practices on available habitat.

Objectives of this research have been 1) to integrate techniques of multivariate classification with the predictive ability of a forest stand simulation model, and 2) to use the resulting model to determine effects of forest management practices on availability of nongame bird habitat. This synthesis requires: a) a structural classification of forest stands in terms of suitability for specific bird species, and b) ability of the forest stand simulator to generate specific variables on which the classification is based. By introducing disturbances (e.g., fire, timber harvest) to the model, we can evaluate effects of natural and man-induced perturbations on availability of habitat for a specific species of bird.

The remainder of this paper will be devoted to presenting an example of this method, FORHAB, a forest stand simulation model designed to predict impacts of certain forest management decisions on availability of habitat for the avian community inhabiting the Walker Branch Watershed.

The Walker Branch Watershed is a 97.5 ha site on the D.O.E. reservation in Anderson County, Tennessee (fig. 2). The watershed ranges in elevation from 285 m to 375 m and occupies an area of steeply sloping ridges and narrow valleys. Girgal and Goldstein (1971), in an analysis of the structure of the watershed, found dominant forest types to be pine (predominately Pinus echinata), yellow popular (dominated by Liriodendron tulipifera), oak-hickory (mixed Quercus spp., Carya spp.) and chestnut oak (typified by Quercus prinus).

MODEL

The following model, FORHAB, is a modified version of FORET (Shugart and West 1977), an Appalachian deciduous forest stand simulator. A detailed description of the model (FORET) can be found in Shugart and West (1977). For the purpose of this paper we will review briefly the general form of the model, but will discuss only modifications in detail.

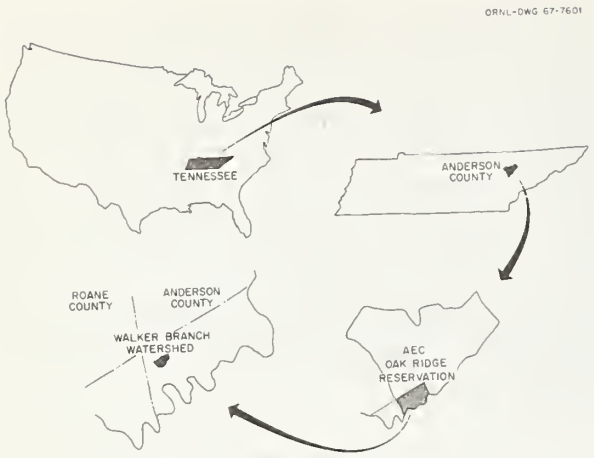


Figure 2. Location of Walker Branch Watershed, Oak Ridge, Tennessee.

FORHAB, like its parent model, simulates the annual change of a forest stand (0.085 ha circular plot) by calculating the growth increment of each tree growing on the stand (subroutine GROW), by tabulating the addition of new saplings to the stand (subroutines BIRTH and SPROUT), and by tabulating the death of trees present on the stand (subroutine KILL). These processes are modeled as stochastic functions. Growth parameters for tree species and climatic conditions included in the model are based on sites of lower slope positions located in eastern Tennessee. A flow chart for FORHAB is provided in appendix I. The main equations for the above subroutines are summarized in appendix II. CUT, HABIT and DISCRM subroutines will be dealt with separately.

Subroutine CUT

The CUT subroutine simulates various forest management practices which are applicable to the southeastern deciduous forest type. The version of this subroutine used for the following analysis was a diameter-limit cut. In this subroutine all commercially valuable species for sawtimber greater than 23 cm dbh (diameter at breast height) were removed from the plot on a 60-year rotation. This form of timber management was practiced on the watershed prior to 1940. The rotation period of 60 years was determined by analysis of stem and basal area curves generated by FORHAB after initial simulations of logging on the watershed.

Subroutine HABIT

The process of classifying stands by their potential to provide habitat for a given bird species is carried out in subroutine DISCRM. The classification is based the following biomass variables which describe vegetational structure of the forest stand:

Foliage biomass of trees 1.2-8.4 cm dbh
Foliage biomass of trees 8.5-22.8 cm dbh
Foliage biomass of trees >22.8 cm dbh
Branch biomass of trees 1.2-8.4 cm dbh
Branch biomass of trees 8.5-22.8 cm dbh
Branch biomass of trees >22.8 cm dbh
Bole biomass of trees 1.2-8.4 cm dbh
Bole biomass of trees 8.5-22.8 cm dbh
Bole biomass of trees >22.8 cm dbh
Number of trees 1.2-8.4 cm dbh
Number of trees 8.5-22.8 cm dbh
Number of trees >22.8 cm dbh
Foliage biomass of average tree 1.2-8.4 cm dbh
Foliage biomass of average tree 8.5-22.8 cm dbh
Foliage biomass of average tree >22.8 cm dbh
Branch biomass of average tree 1.2-8.4 cm dbh
Branch biomass of average tree 8.5-22.8 cm dbh
Branch biomass of average tree >22.8 cm dbh
Bole biomass of average tree 1.2-8.4 cm dbh
Bole biomass of average tree 8.5-22.8 cm dbh
Bole biomass of average tree >22.8 cm dbh

Model output in the form of species and tree diameters must be used to calculate these biomass variables.

The HABIT subroutine divides all trees on the simulated plot into two groups, conifer and deciduous, and then into three size classes within each of these two groups. The foliage, branch and bole biomass for each tree is then calculated using regression equations in table 1, which are site specific to the Walker Branch Watershed (Harris et al. 1973). These values are then summed for all trees on a plot for each size class to provide the variables listed above.

Subroutine DISCRM

The classification of simulated forest stands as potential habitat for a given bird species is carried out in subroutine DISCRM. The classification is based on the statistical procedure of two-group discriminant function analysis (Morrison 1967). Classification criteria were constructed using vegetation data collected on 298 0.085-ha permanent census plots. Breeding territories of various bird species were located and mapped if they either contained or overlapped any of the 298 plots. If a plot was located within the territory of an individual bird (breeding pair), that plot was considered as potential habitat for that species. Conversely, if a given plot was not within the boundary of a territory of an individual of that species, that plot was classified as not providing habitat for that species. Thus data were obtained on areas of both suitable habitat and areas considered inadequate for the needs of the various species. Data on vegetation of these census plots, in the form of species and diameter for each tree on the plot, then were used to generate biomass variables for classification using the same regression equations as those in subroutine HABIT.

Table 1. Uncorrected regressions (ln-ln) of tree component weight (kg) on tree diameter breast height (cm) used in subroutine HABIT.

Dependent variable		Intercept (a)	Slope (b)	R^2	N[1]	K[2]
Leaf	All spp.	-3.498	1.695	0.86	302	1.34
	Hardwoods	-3.862	1.740	0.88	178	
	Conifers	-2.907	1.674	0.91	65	
Branch	All spp.	-3.188	2.226	0.91	298	1.26
	Hardwoods	-3.173	2.224	0.89	231	
	Conifers	-3.461	2.292	0.95	51	
Bole	All spp.	-2.437	2.418	0.97	298	1.08
	Hardwoods	-2.270	2.385	0.98	231	
	Conifers	-3.787	2.767	0.96	51	
Branch-bole	All spp.	-2.126	2.393	0.96	371	

[1]N = Number of samples in each regression
[2]K is the correction factor for bias on logarithmic transformation which is multiplied by exp [ln $\hat{Y}$]

Figure 3 is a hypothetical example of the actual classification procedure in subroutine DISCRM. For each bird species, census plots were placed into one of two groups, suitable habitat or unsuitable habitat, based on the process described above. The two elipses in figure 3 represent two populations of census plots plotted in two-dimensional space (with X1 and X2 being two biomass variables from table 1). These two populations were then subjected to two group discriminant function analysis.

Discriminant function analysis is a statistical procedure for finding a linear combination of the original predictor variables (X1 and X2) which results in the largest difference between the two mean vectors (i.e., maximizes the ratio of among-group to within-group sums of squares). Let us examine the case of two populations, suitable habitat and unsuitable habitat plots, with samples of N1 and N2 independent observations. The populations of P (number of variables) responses are multivariate normal with a common variance-covariance matrix Σ, but different mean vectors μ and μ_2. If $\underline{X}1$ and $\underline{X}2$ are the sample mean vectors for the two groups and S is the pooled estimate of Σ, our intention is to find a coefficient vector $\underline{a}$ of the linear compound $\underline{a}X$ of the responses which will maximize the distance between the two groups in P-dimensional space. It can be shown that

$$\underline{a} = S^{-1}(\underline{X}1 - \underline{X}2);$$

and the general form of the resulting discriminant function is

$$Y = (\underline{X}1 - \underline{X}2)'S^{-1}X$$

with

$$Y_i = a_1X_1 + a_2X_2 + \ldots + a_pX_p$$

(where p is the number of variables in the classification).

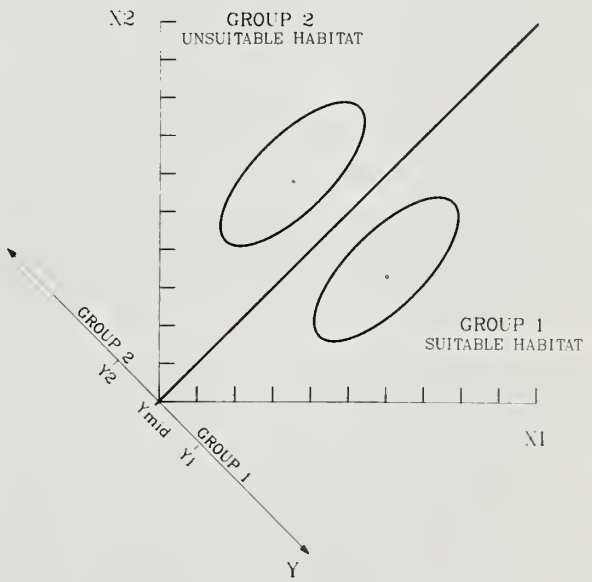


Figure 3. Linear decision scale from subroutine DISCRIM.

In the case of two populations with multivariate normal distributions and respective mean vectors $\underline{X}_1$ and $\underline{X}_2$, common covariance matrix S and prior probabilities of group membership h and (1-h), the Bayes, or minimum expected loss, classification rule states that the observation vector $\underline{X}$ should be assigned to population 1 if

$$\underline{X}'S^{-1}(\bar{\underline{X}}_1-\bar{\underline{X}}_2) - 1/2(\bar{\underline{X}}_1+\bar{\underline{X}}_2)'S^{-1}(\bar{\underline{X}}_1-\bar{\underline{X}}_2) \geq \ln \frac{(1-h)}{h}$$

and assigned to population 2 if this relationship does not hold.

It should be noted that the first term in this equation is the linear discriminant score for the observation vector $\underline{X}$ and the second term is the point midway between the means of the discriminant function as computed for each group or population. For the purpose of classifying simulated observation vectors (subroutine HABIT) the prior probabilities of group membership can be assigned to be equal since there is no a priori reason for assuming group membership. As a result of this assumption, the term to the right of the inequality becomes zero and the equation can be expressed as

$$\underline{X}'S^{-1}(\bar{\underline{X}}_1-\bar{\underline{X}}_2) \geq 1/2(\bar{\underline{X}}_1+\bar{\underline{X}}_2)'S^{-1}(\bar{\underline{X}}_1-\bar{\underline{X}}_2)$$

with the term to the left of the inequality being the discriminant score for the observation vector to be classified and term to the right being the midpoint between the discriminant scores as computed for the group mean vectors.

By calculating the Y_i scores corresponding to the mean vectors $\bar{\underline{X}}1$ and $\bar{\underline{X}}2$ (where X1 = suitable habitat and X2 = unsuitable habitat) and then taking the midpoint of those values;

$$Y_{mid} = (Y1 + Y2)/ 2$$

one can compute the linear decision scale as shown in figure 3.

To classify a newly sampled response (such as a plot output by the model) into either group 1 or 2, the Y_i score must be calculated for that given response vector. If the Y_i value is to the group 1 direction of Y_{mid} then the plot would be classified as belonging to group 1, with the converse holding true if the Y_i value was to the opposite direction of Y_{mid}.

Subroutine DISCRM consists of a series of these linear decision scales, one corresponding to each bird species comprising the avian community on the watershed. Each simulated plot is input to subroutine HABIT where the biomass variables necessary for classification are generated. These variables are then input to subroutine DISCRM where the Y_i value is calculated for that plot. This value is then compared to the Y_{mid} value for each bird species and the decision as to whether that plot provides potential breeding habitat for each of the species of interest is output from the model.

RESULTS

A test of homogeneity of within covariance matrices for the two group discriminant function analysis of ovenbird (Seiurus aurocapillus) habitat found no significant difference between covariance matrices (as approximated by χ^2, $P \geq 0.05$) and they were pooled for the analysis (Kendall and Stuart 1961). The two groups (suitable and unsuitable habitat) were found to be significantly different ($P \geq 0.05$) with respect to those structural variables measured, using the Hotelling T^2 test statistic. Bayesian classification of initial observations (298 plots) based on the two group discriminant function analysis accurately classified group membership (suitable and unsuitable habitat) for 96% of the forested plots, with prior probabilities of group membership being set proportional to the number of observations in each group.

Figure 4 shows results of a 500-year simulation of available habitat for the ovenbird (Seiurus aurocapillus) on Walker Branch Watershed. Results are presented as percentage of total land area which provides potential breeding habitat for the ovenbird over a 500-year period. Year zero represents the present structural configuration of the forest on the watershed. This was accomplished by initializing the model with 25 randomly chosen census plots from the watershed using vegetation data collected on these plots in 1977. Results are given for both simulations including timber management (diameter-limit cut) and undisturbed conditions. Simulations of undisturbed forest dynamics show an initial increase in available habitat for the ovenbird from 20 percent of the land area to approximately 65 percent over the next 60 years. This decline to year 90 is followed by a general oscillation of available habitat from 10 to 20 percent for the remainder of the simulation with only one period of increase above 30 percent, that being at approximately year 250.

In comparison, results of the simulation which included timber management showed considerable divergence from simulations of the undisturbed forest. Dynamics previous to the first cut at year 60 were identical for both managed and undisturbed simulations. Following the first cut at year 60, however, we see a divergence with the managed stands showing an increase in available habitat to 85 percent by year 90 as compared to the 5 percent available habitat for the undisturbed simulations. This increase is followed by a continuous decline over the next 30 years until the second cut at year 120. At this time once again we see an increase in available habitat for the ovenbird to 50

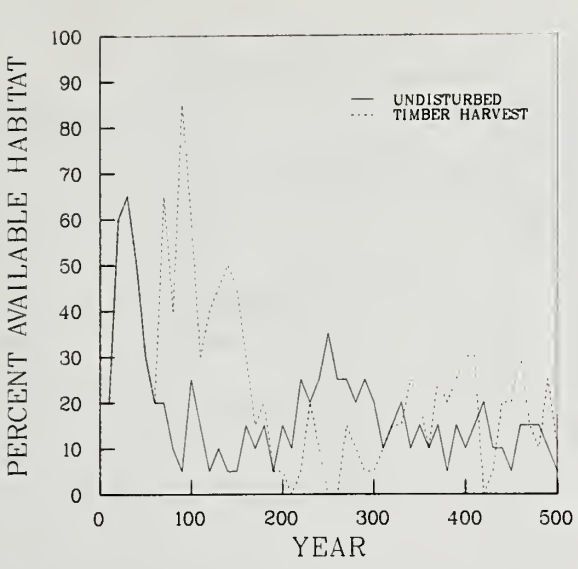


Figure 4. Results from 500-year simulation of ovenbird habitat on the Walker Branch Watershed in east Tennessee. Available habitat expressed as percentage of total land area of the watershed.

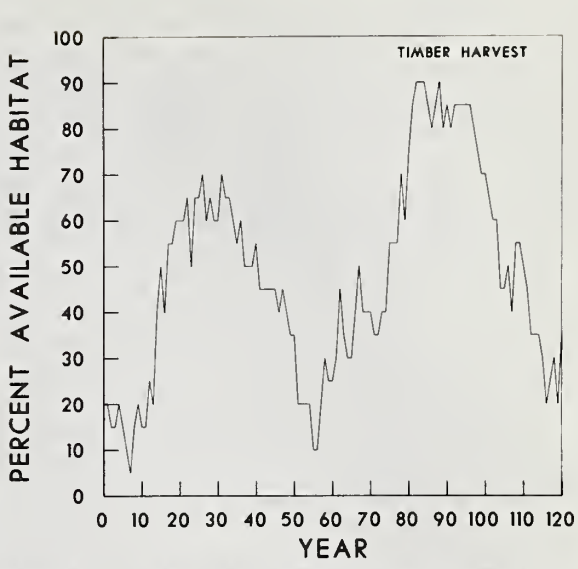


Figure 5. Results from 120 year simulation of available ovenbird habitat on the Walker Branch Watershed under conditions of timber management. Available habitat expressed as percentages of the total land area of the watershed.

percent of the total forested area of the watershed. Following the second timber harvest at year 120 there is a decline in available habitat continuing to the next cut at year 180. Following the third cut there is a continuing decline to year 210, when the model predicts there will be no potential breeding habitat for the ovenbird on the watershed. In all remaining cuts there is an initial increase in available habitat following harvest. This is followed by a decline, which generally continues until the next cut, the last being at year 480.

Figure 5 presents results of the same simulation under conditions of timber management that were presented in figure 4, only over a shorter time scale showing the short-term dynamics of available habitat following timber harvest (year 60). Once again there is an initial increase in availability of suitable habitat followed by a decline at year 30. This decline continues until the forest is cut at year 60, when those trees greater than 23 cm dbh and of commercial value as sawtimber are removed from the forest. This thinning of the forest results in an increase of potential habitat for the ovenbird over the next 35 years, at which time a downward trend begins and continues through the remainder of the simulation.

DISCUSSION

Results presented in Figures 4 and 5 represent predicted dynamics of available habitat for the ovenbird on the Walker Branch Watershed. The 500-year simulation in figure 5 is meant to show that extrapolations cannot necessarily be made from results of a single timber cut to cuts pending in the future. Resulting availability of habitat for the ovenbird following an initial diameter-limit cut on a 60-year rotation may not represent availability of potential habitat the forest will provide following future diameter-limit cuts. Availability of habitat following the first cut increased dramatically, but subsequent cuts on a 60-year rotation yielded quite different results. Secondly, failure to look at potential effects of repeated harvests may mislead the manager with respect to long term habitat dynamics. As can be seen in results of the 500-year simulation, the first cut was followed by a four-fold increase in available habitat for the ovenbird. Subsequent cuts, however, resulted in less dramatic increases and in some cases led to elimination of potential habitat.

These results show the importance of historic considerations in determining effects of particular timber management practice on a given forest. Structural configuration of the forest prior to cutting is of utmost importance in the case of repeated long-term management plans. To date this type of information has been lacking. FORHAB and models of its type can be used to provide information on long-term management plans and combinations of management schemes before their actual implementation.

Results (figs. 4 and 5) show an initial

increase in potential habitat after the initial cut at 60 years. This increase contrasts with decline of habitat availability for undisturbed simulations. The increase after cutting is a result of thinning or general decrease in density of stands on the watershed. At present the forest has gone approximately 70 years without a timber harvest, so by year 60 the stand is approximately 130 years in age. Thinning of larger trees, which leads to an initial increase in available habitat, is quickly followed by an increased density of the understory, reducing available habitat. This is the reason for the pattern of initial increase after cutting followed by a subsequent decline in availability of habitat for the ovenbird. These results are in general agreement with the structure of vegetation chosen as breeding habitat by the ovenbird on the watershed. Optimal habitat for the ovenbird on the Walker Branch Watershed (as determined by the relationship between vegetational structure and prey abundance) is a deciduous stand with a sparse understory and brush and little ground cover.[5]

It should be noted that the model simulates availability of potential breeding habitat expressed as a percentage of the total area under consideration. The model does not simulate population dynamics of a given bird species per se. The ability of the ovenbird population to track changes in availability of habitat such as the initial increase in habitat following the first timber harvest, or to reinvade after the disappearance of potential breeding habitat in an area are not considered explicitly in the model. These considerations would depend on immigration into the area or the existence of a "floater" population of nonbreeding individuals unable to establish territories due to lack of suitable habitat. Likewise, the model does not consider quality of habitat provided by a given stand. Some marginal areas may become potential habitat depending on size of the ovenbird population. These points must be kept in mind when interpreting results of simulation from the model.

The potential of predicting effects of proposed management schemes on habitat availability need not be limited to the breeding season during spring and summer months. For many resident bird species, availability of suitable habitat during fall and winter seasons may be the major limiting factor. Figure 6 shows seasonal variation in habitat selection for several species of birds on the watershed as expressed by the first two discriminant functions describing the vegetational structure of the forest stands (Shugart et al. 1975). It can be seen that in many cases the habitat needs of certain species change seasonally. By including information on seasonal habitat requirements into the classification procedure, FORHAB and models of its type could be used to determine habitat availability not only on a yearly basis, but

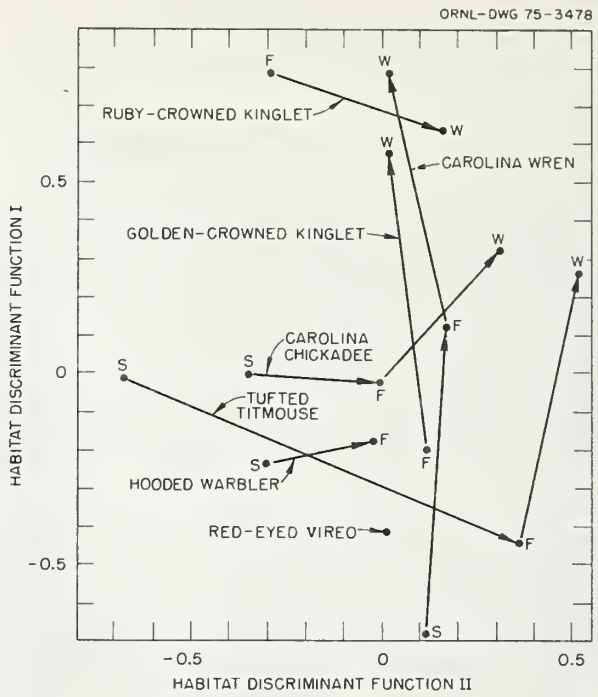


Figure 6. Seasonal variation in habitat selection for several species of birds inhabiting the Walker Branch Watershed.

seasonal variations in available habitat.

Two-group discriminant function analysis (as described in subroutine DISCRM) can be used to construct habitat maps which reflect the potential of a site to provide habitat for a given species of bird. A map of potential habitat for five woodpecker species on the Haw Ridge Watershed on the D.O.E. reservation in east Tennessee is shown in figure 7 (Shugart et al. 1978). The map was constructed using a classification scheme identical to that presented for subroutine DISCRM. By initializing the model for a given forested region such as the Haw Ridge Watershed (initial conditions based on sample plots from the Haw Ridge Watershed as was done for the Walker Branch Watershed) it would be possible to construct a series of maps representing changes in available habitat for the area through time. This would aid in more site specific, small scale management problems.

The model FORHAB which has been presented as an example of the process and methodology of habitat simulation has dealt solely with the Appalachian deciduous forest of the Southeast. However, the general form of the forest simulation model underlying this method (FORET) has been adapted for a wide variety of forested areas, including mixed conifer-hardwood forests of the Northeast (Botkin et al. 1972), loblolly pine forests of Arkansas (Milke 1978) and rain forests of Puerto Rico (Doyle 1980). Another version of

[5]T.M. Smith, unpublished data on file at Oak Ridge National Laboratory.

Figure 7. Map of potential habitat for five species of woodpecker on Haw Ridge as defined by discriminant function classification procedure.

the model FORMIS[6] simulates flood plain forests of the Mississippi River and could be modified to simulate riparian habitats in other areas.

Potential applications of the model to predict effects of proposed and untried management schemes on specific forested areas for both individual species and the avian community as a whole, as well as its versatility and adaptability to a diverse array of forested regions, make models like FORHAB potentially important tools for forest and wildlife managers in the future.

ACKNOWLEDGMENTS

Research supported by National Science Foundation under Interagency Agreement 40-700-78 with U.S. Department of Energy under contract W-7405-eng-26 with Union Carbide Corporation.

LITERATURE CITED

Anderson, S.H., and H.H. Shugart. 1974. Habitat selection of breeding birds in an East Tennessee deciduous forest. Ecology 55:828-837.

Botkin, D.B., Jr., J.F. Janek, and J.R. Wallis. 1972. Some consequences of a computer model of forest growth. Journal of Ecology 60:849-872.

Conner, R.N., and C.S. Adkisson. 1976. Discriminant function analysis: A possible aid in determining the impact of forest management on woodpecker nesting habitat. Forest Science 22:122-127.

DeGraaf, R.M., technical coordinator. 1978. Workshop on management of southern forests for nongame birds. USDA Forest Service General Technical Report SE-14, 175 p. Southeastern Forest Experiment Station, Asheville, N. C.

Doyle, T.W. 1980. FORICO: Gap dynamics model of the lower montane rain forest in Puerto Rico. 94 p. M.S. Thesis, University of Tennessee, Knoxville.

Franzreb, K.E., and R.D. Ohmart. 1978. The effects of timber harvest on breeding birds in a mixed coniferous forest. Condor 80:431-441.

Grigal, D.F., and R.A. Goldstein. 1971. An integrated ordination-classification analysis of an intensively sampled oak-hickory forest. Journal of Ecology 59:481-492.

Hagar, D.C. 1960. The interrelationships of logging, birds, and timber regeneration in the douglas-fir region of northwest California. Ecology 41(1):116-125.

Harris, W.F., R.A. Goldstein, and P. Sollins. 1973. Net above ground production and estimates of standing biomass on Walker Branch Watershed. p. 41-64. In Young, H.E., editor. IUFRO Biomass Studies. University of Maine Press, Orono, Me.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Kasanaga, H., and M. Mousi. 1954. On the light transmission of leaves and its meaning for the production of matter in plant communities. Japanese Journal of Botany 14:302-324.

Kendall, M.G., and A. Stuart. 1961. The advance theory of statistics. p. 266-282. Volume 3. Charles Griffin and Co., London.

Ker, J.W., and J.H.G. Smith. 1955. Advantages of the parabolic expression of height-diameter relationships. Forest Chronology 31:235-246.

Loomis, R.S., W.A. Williams, and W.G. Duncan. 1967. Community architecture and the productivity of terrestial plant communities. p. 291-308. In San Pietro, A., F.A. Greer, and T.J. Army, editors. Harvesting the sun. Academic Press, New York, N.Y.

MacArthur, R.H., and J.W. MacArthur. 1961. On bird species diversity. Ecology 42:594-598.

[6]M.L. Tharp, unpublished model available at Oak Ridge National Laboratory.

Milke, D.L., H.H. Shugart, and D.C. West. 1978. Users manual for FORAR, a stand model for composition and growth of upland forests of southern Arkansas. ORNL/TM-5767, ESD Publication No. 1021. Oak Ridge National Laboratory, Oak Ridge, Tenn.

Morrison, D.F. 1967. Multivariate statistical methods. 415 p. McGraw Hill, New York, N.Y.

Perry, T.O., H.E. Sellers, and C.O. Blanchard. 1969. Estimation of photo-synthetically active radiation under a forest canopy with chlorophyll extracts and from basal area measurements. Ecology 50:39-44.

Shugart, H.H., S.H. Anderson, and R.H. Strand. 1975. Dominant patterns in bird populations of the eastern deciduous forest biome. p. 90-95. In Smith, D.R., techincal coordinator. Management of forest and range habitats for nongame birds: Proceedings of a symposium [Tucson, Ariz., May 6-7, 1975]. USDA Forest Service General Technical Report WO-1, 343 p. Washington, D.C.

Shugart, H.H., and B.C. Patten. 1972. Niche quantification and the concept of niche pattern. p. 284-326. In Patten, B.C., editor. Systems analysis and simulation in ecology II. Academic Press, New York, N.Y.

Shugart, H.H., T.M. Smith, J.T. Kitchings, and R.L. Kroodsma. 1978. The relationship of nongame birds to southern forest types and successional stages. p. 5-16. In DeGraaf, R.M., techincal coordinator. Management of southern forests for nongame birds: Proceedings of a workshop [Atlanta, Ga., January 24-26, 1978] USDA Forest Service General Technical Report SE-14, 175 p. Southeastern Forest Experiment Station, Asheville, N. C.

Shugart, H.H., and D.C. West. 1977. Development of an Appalachian deciduous forest succession model and its application to assessment of the impact of the chestnut blight. Journal of Environmental Management 5:161-179.

Shugart, H.H., and D.C. West. 1980. Forest succession models. Bioscience 30(5):308-313.

Slusher, J.P., and T.M. Hinckley. 1974. Timber-wildlife management: Proceedings of a symposium. [Columbia, Mo., January 22-24, 1974] 131 p. Missouri Academy of Science. Occasional Paper No. 3, Columbia, Mo.

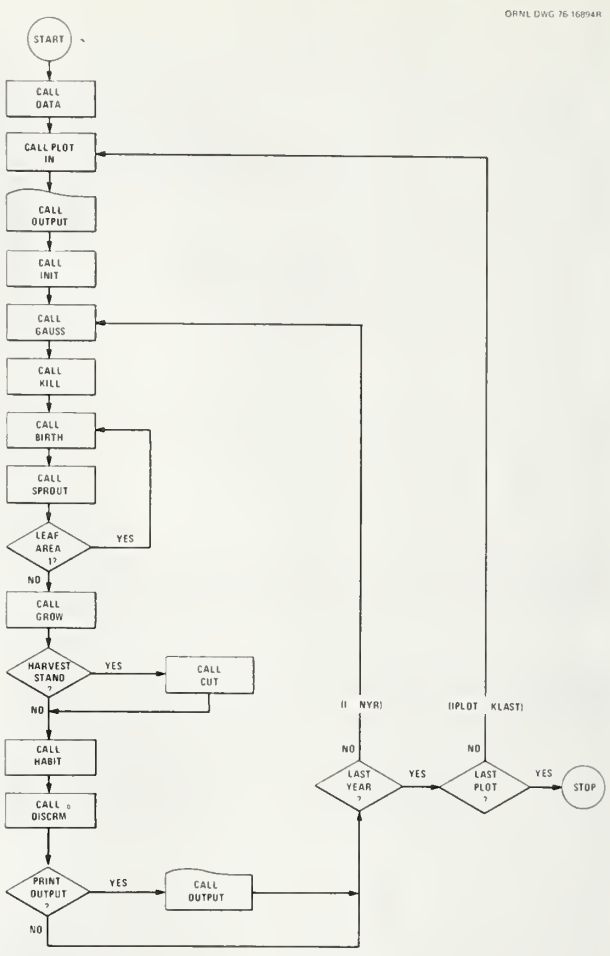
Smith, D.R., technical coordinator. 1975. Management of forest and range habitats for nongame birds. Proceedings of a smyposium [Tucson, Ariz., May 6-9, 1975]. USDA Forest Service General Technical Report WO-1, 343 p. Washington, D.C.

Smith, R.L. 1980. Ecology and field biology. Third edition. 835 p. Harper and Row, New York, N.Y.

Thomas, J.W., editor. 1979. Wildlife habitats in managed forests the Blue mountains of Oregon and Washington. 512 p. USDA Forest Service, Agriculture Handbook 553, Washington, D.C.

Whitmore, R.C. 1975. Habitat ordination of passerine birds of the Virgin River Valley, southwestern Utah. Wilson Bulletin 87:65-74.

Appendix I. Flow diagram for FORHAB.



Appendix II. Model equations in FORHAB.

Process	Equations
Growth of each tree under optimal conditions[1]	$\frac{d[D^2H]}{dt} = R\ LA\ (1 - \frac{DH}{D_{max}H_{max}})$ R = growth rate parameter, LA = leaf area of tree, D = diameter at breast height, H = height of tree, D_{max} = maximum diameter for a particular species, H_{max} = maximum height for a particular species, D^2H = index of tree volume.
Height/diameter relation[2]	$H - 137 + b_2D - b_3D^2$, $b_2 = 2(H_{max} - 137)/D_{max}$, $b_3 = (M_{max} - 137)/D^2_{max}$, b_2 and b_3 determined by setting $H = H_{max}$ and $dH/dt = 0$ when $D = D_{max}$.
Extinction of light as a function of leaf area in forest canopies[3]	$Q(h) = Q_o \exp(-k \int_h^\infty LA(h')dh')$ LA(h') = distribution of leaf area as a function of height Q_o = incident radiation, Q(h) = radiation at height (h), K = constant[3].
Reduction of photosynthesis due to shading[4]	Various empirical equations fitted to light-photosynthesis curves for shade-tolerant or shade-intolerant species found in each forest. These equations are used to reduce the magnitude of the growth equation (above) for shaded trees.
Crowding effects related to stand biomass (competition)[1]	S(BAR) = 1 - BAR/SDILQ, BAR = total biomass (basal area) of simulated stand, SDILQ = maximum biomass (basal area) recorded.
Intrinsic tree mortality[1]	$p = 1 - (1 - \epsilon)^n$, p = probability of mortality at year n is chosen such that p = D.99 when $n = AGE_{max}$ (the maximum age for the species).
Mortality of trees with suppressed growth[1]	If growth is less than a critical value for the species, p = 0.368.

[1]Botkin et al. 1972.

[2]Ker and Smith 1955.

[3]Kasanaga and Monsi 1954, Loomis et al. 1967, Perry et al. 1969.

[4]Kramer and Kozlowski 196D.

A WINDSHIELD AND MULTIVARIATE APPROACH TO THE CLASSIFICATION, INVENTORY, AND EVALUATION OF WILDLIFE HABITAT: AN EXPLORATORY STUDY[1]

C.E. Grue[2], R.R. Reid[3], and N.J. Silvy[4]

Abstract.--Techniques are described for evaluating the habitats of breeding mourning dove (Zenaida macroura), and bobwhite (Colinus virginianus) and scaled quail (Callipepla squamata) from within a vehicle. Audio counts of the three species and habitat surveys were conducted on 133 (24 km) transects in Texas in 1976. The linear distance of each habitat type intersecting a transect and the number of structural features within ca. 0.8 km were recorded. Habitat variables, including indices of habitat interspersion and diversity, were analyzed for correlations with audio counts of the three species using stepwise multiple regression. We used discriminant analysis to determine the accuracy of using habitat variables to identify transects supporting below or above average densities of the three species.

Habitat variables accounted for 28 to 87% of the variation in mourning dove call counts, and 60 to 95% and 32 to 93% of the variation in bobwhite and scaled quail whistle counts, respectively. Discriminant analyses correctly classified 78 to 89% of the mourning dove call-count surveys. Comparable values for whistle counts of bobwhite and scaled quail were 71 to 96% and 39 to 98%, respectively. Results suggest techniques developed may be applicable to analyzing the habitat of wildlife species for which transects are used to obtain population estimates.

Key words: Bobwhite quail; discriminant analysis; habitat classification; habitat evaluation; habitat inventory; mourning dove; multiple regression; scaled quail; Texas; windshield approach.

[1]Texas Agricultural Experiment Station Technical Article 16235. Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Graduate Research Assistant, Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX 77843. Present position: Research Biologist, U.S. Fish and Wildlife Service, Patuxent Wildlife Research Center, Laurel, MD 20811.

[3]Graduate Research Assistant, Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX 77843. Present position: Staff Biologist II, Espey-Huston and Associates, Inc., P.O. Box 519, Austin, TX 78767.

[4]Associate Professor, Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX 77843. Present position: Assistant Leader, Florida Cooperative Fish and Wildlife Research Unit, 117 Newins-Ziegler Hall, University of Florida, Gainesville, FL 32611.

INTRODUCTION

The abundance of wildlife is usually determined by habitat conditions which alter the carrying capacity of the land. These conditions do not remain static. Our understanding of trends in habitat quality and quantity is limited, even for those wildlife species for which habitat needs are generally known. It is, therefore, essential to determine the habitat requirements of wildlife, develop the capability to assess the abundance of critical habitats, and monitor changes in their quantity and quality.

The mourning dove is a species which has not received management commensurate with its popularity as a game bird (Amend 1969, Sandfort 1977). This species has the widest range of any game bird in the United States and is the most important in North America in terms of numbers and hunter harvest (Keeler 1977). Management of mourning doves in the United States, however, has been almost entirely restricted to control of harvest based on fluctuations in breeding populations monitored nationwide by call-count surveys (Dolton 1977). While some breeding populations have decreased (Dolton 1977), hunter harvest has increased to more than 49 million (Keeler 1977). If demand for utilization of the dove resource continues to increase, habitat management may become essential. First steps in initiating management plans will be habitat inventory, analysis, and evaluation. The objective of our study was to develop techniques for evaluating the habitat of breeding mourning dove from a vehicle ("windshield approach"). Texas was well suited for such a study because of its size and habitat diversity (Gould 1975). In addition, we concurrently evaluated habitats of breeding bobwhite and scaled quail using these same techniques to determine if they were applicable to other wildlife species.

METHODS

Our study of habitats of breeding mourning dove and bobwhite and scaled quail in Texas consisted of four steps: classification, inventory, analysis, and evaluation. Classification was defined as identification and placement of habitats into specific habitat types according to established criteria. Inventory was defined as the process by which abundance of a particular wildlife species or habitat parameter was determined along an audio-count transect. We defined analysis as examination of habitat, its types and structural features, and their relationships to audio counts and one another. Evaluation was defined as the process by which habitats were ranked according to estimated densities of the three species they supported. From these data, use of habitat variables to predict audio counts of the three species was examined.

Classification

We developed a method of classifying habitats from within a vehicle. Habitat type was defined as a description of the vegetation of an area consisting of a unique combination of canopy composition and spatial distribution and ground cover height and composition. Our hierarchial habitat classification (fig. 1) was divided horizontally into three major strata. (Figures and tables follow literature cited). The first, "physiognomic class", was used to describe vertical and horizontal distribution of canopy vegetation within a given area (for detailed descriptions, see Grue 1977). The second level subdivided habitats containing trees on the basis of canopy composition (deciduous, coniferous, or mixed) and presence or absence of understory. In the present study, we also separated mesquite (Prosopis spp.) from other deciduous species. In the third level, cropland, pasture, savannah, parkland, desert scrub, woodland, and forest were further divided based on height and/or composition of ground cover. Classification of canopy and ground cover composition and ground cover height was based on relative abundance within each composition and height category. If at least 75% of canopy or ground cover was similar in composition and height, it was considered homogeneous. Canopy or ground cover in which less than 75% was similar in composition was considered mixed.

We also considered structural features within habitat types; structures or characteristics other than height and composition of ground cover, and composition and spatial distribution of canopy, which others (for review, see Grue 1977, Reid 1977) have suggested may be important to breeding dove and quail. Included within this category were the number of fences, shrubrows, windbreaks, powerlines, roads, and railroad rights-of-way, and whether or not these structures paralleled or intersected the call-count transects. The number of edges (an abrupt change in the physiognomy of the vegetation excluding ecotones), permanent water sources, buildings with associated vegetation, washes, livestock feeders and feedlots, gravel pits, irrigation and oil pumps, and presence of snags (dead, defoliated woody shrubs or trees) within 0.8 km of each transect were included. Type of road surface on the survey route (asphalt, gravel, sand, or dirt) and the width of the road shoulder were recorded at each stop.

Inventory

Call-count and whistle-count data for 1976 were obtained for the first 15 (3-min) listening points (stops) located at 1.6 km intervals along each of the 133 Federal and State mourning dove call-count transects in Texas. Call counts were conducted on each transect four times between 20 May and 10 June 1976 by personnel of the Texas Parks and Wildlife Department (Dunks 1976). Quail

whistle counts were conducted concurrently on the last three of the four surveys of each transect. Audio-counts are believed to be reliable indices of the relative abundance of breeding dove (Dolton 1977) and quail (Bennitt 1951, Elder 1956, Rosene 1957, Norton et al. 1961, Campbell et al. 1973, Brown et al. 1978) within relatively large areas.

The habitat intersecting the transects was surveyed during this 20-day time period using two vehicles, each with a two-man team (Grue et al. 1976). To reduce error due to differences in observers, trial surveys were conducted as a group along several transects throughout most of Texas; each team worked within different ecological areas (Gould 1975) and one member of each team was designated driver and odometer reader, while the other person classified habitat throughout the study. A team traveled and recorded habitat data on both sides of each transect starting 0.8 km before and ending 0.8 km after each stop. Each of these 1.6 km units was defined as a transect interval. The linear distance of each observation of a habitat type intersecting the survey route, measured to the nearest 0.02 km, and the number of structural features present within 0.8 km (maximum radius of audibility of each species; Davey 1955, Baxter and Wolfe 1973) were also recorded within each transect interval.

Analysis and Evaluation

Habitat variables were analyzed for correlation with audio counts of the three species by transect, statewide, and within ecological areas, using stepwise multiple regression (Barr and Goodnight 1972). Ecological areas of Gould (1975) were selected because call counts were more homogeneous within their boundaries than those of other physiographic divisions of the State (Grue 1977) and have been used by the Texas Parks and Wildlife Department to analyze annual call-count data (Dunks 1976). Call-count and whistle-count data for all surveys considered valid by the Texas Parks and Wildlife Department were included in analyses because variation in audio counts of the three species between surveys was significant (Grue 1977, Reid 1977). Independent variables (habitat variables) entered and remained in models if values for their partial F-statistics were significant ($P<0.05$).

Habitat interspersion and diversity indices were included in all analyses. Interspersion was calculated by summing frequencies of occurrence for each habitat type within the 15 transect intervals. Habitat diversity was calculated for each transect using the Shannon-Weiner Index (Shannon 1948). Transect call counts and whistle counts were equal to the sum of the number of dove and quail heard calling on the 15 stops, respectively.

To determine importance of spatial distribution of the canopy, canopy composition, and ground cover height and composition in predicting audio counts of the three species, we evaluated seven simplifications of our habitat classification. Simplifications corresponded to horizontal strata within the hierarchy of the initial classification from the most complex (habitat type = physiognomic class + canopy composition + ground cover height and composition) to the least complex (habitat type = physiognomic class). Simplifications were evaluated for each of the three species by ecological area using transect audio counts and habitat variables. Structural features were excluded from stepwise multiple regression analyses so that multiple correlation coefficients represented differences between habitat classifications. Indices for habitat interspersion and diversity were also not included because of the number of simplifications examined. The simplified habitat classification with the fewest habitat variables which also resulted in high multiple correlation coefficients statewide was judged the "best" classification system for each species.

We developed an index to minimum habitat interspersion applicable to the simplified habitat classifications selected. Habitat interspersion was more difficult to calculate than habitat diversity (we continued to use the Shannon-Wiener Index) because the number of times the habitat types actually changed on both sides of a call-count transect was not known. The new interspersion index was based on the number of habitat types present within a transect as well as the presence or absence of each habitat type within adjacent transect intervals. If a particular habitat type was present within a transect interval but was absent within an adjacent interval, the value of the interspersion index increased by 1. Conversely, if a particular habitat type was present or absent within two adjacent transect intervals, interspersion was equal to 0 and the value of the index remained unchanged. This process was continued until all habitat types were examined within the 15 transect intervals of each of the 133 call-count transects. The interspersion index on a particular transect was equal to this value plus the number of habitat types present within the transect. The latter was used as an indirect measure of minimum interspersion within the transect intervals.

We used discriminant analysis (Barr and Goodnight 1972) to determine the accuracy of predicting audio-count classes of the three species using the habitat variables within the simplified habitat classifications selected. Analyses were conducted by ecological area and only those habitat parameters within the multiple linear regression models for a given ecological area were included. A mean transect audio-count was determined for each species and ecological area using data from all call-count and whistle-count surveys conducted in 1976 (table 1). Transect audio counts for each species and survey were then classified into one of two classes: average or above, or below average. The use of habitat variables to predict audio-count classes was evaluated in terms of percent of call-count and whistle-count surveys correctly classified

within each class by the discriminant functions. Multiple analysis of variance (Barr and Goodnight 1972) was used to test for statistically significant ($P<0.05$) differences in values for the habitat variables between the two audio-count classes.

RESULTS

Habitat variables within the initial habitat classification accounted for up to 87% of variation in mourning dove call counts, and up to 94 and 93% of variation in bobwhite and scaled quail whistle counts, respectively (table 2). Analyses within ecological areas resulted in an increase in multiple correlation coefficients and/or a decrease in the number of independent variables in the models. Both habitat types and structural features accounted for a significant portion of variation in audio counts of the three species (table 3).

Simplification of the initial habitat classification resulted in a significant reduction in the number of habitat types, but a relatively slight decrease in multiple correlation coefficients (table 4). The number of habitat types within the simplified habitat classification selected for each of the three species was reduced 90 to 96%, while multiple correlation coefficients decreased by only 4 to 14%. Of seven simplified habitat classifications evaluated, physiognomic class with canopy composition and cropland divisions of grain, nongrain, forage, and plowed ground accounted for the most variation in mourning dove call counts. Physiognomic class with canopy composition with mesquite was the simplified habitat classification selected for both species of quail.

Models for predicting audio counts of the three species incorporating both structural features and habitat types within the simplified habitat classifications selected are presented in table 5. Multiple correlation coefficients for these models were similar to those associated with the initial habitat classification (table 2). Models accounted for 28 to 88% of variation in mourning dove call counts, and 60 to 95% and 32 to 93% of variation in bobwhite and scaled quail whistle counts, respectively, within the 10 ecological areas.

Discriminant functions incorporating the habitat variables within these models correctly classified 75 to 89% of mourning dove call-count surveys within each ecological area into the two audio-count classes (table 6). Comparable values for whistle-count surveys of bobwhite and scaled quail were 71 to 96% and 39 to 98%, respectively (table 6). Multiple analyses of variance indicated that there were significant ($P<0.05$) differences between values for the habitat parameters within the discriminant functions for the two audio-count classes for each species.

DISCUSSION

Results suggested that the habitat classification and techniques we employed may identify habitat variables significantly correlated with audio counts of mourning dove and bobwhite and scaled quail. Both habitat types and structural features appeared to be important components of habitats selected by the three species during the breeding season. Analyses by transect within ecological areas usually resulted in an increase in multiple correlation coefficients or a decrease in the number of independent variables in models for the three species. These trends may be explained by increased homogeneity in audio counts (Foote et al. 1958) or habitat within ecological areas (Blankenship et al. 1971). Analysis within ecological areas may also more accurately describe the relationships between audio counts of the three species and habitat variables. Simple correlation analyses (Grue 1977; Reid et al. 1978, 1979) suggested audio counts of the three species were correlated with habitat variables that may have provided requisites necessary for survival and reproduction. Habitat parameters which provided these requisites differed between ecological areas and appeared to depend on the abundance and distribution of the habitat types and structural features present. For example, mourning dove call counts were positively correlated with cropland within the Trans-Pecos, but were negatively correlated with this habitat type on the High Plains. In the Trans-Pecos, nesting substrate was abundant, whereas sources of food and water were generally restricted to cultivated areas. The opposite was true on the High Plains, where food and water were more abundant, but nest sites within woody vegetation were limited. Results suggest future analyses and evaluations should be conducted within physiographic divisions.

With a few exceptions, the amount of variation in audio counts of the three species accounted for by habitat types within the initial and simplified habitat classifications selected was similar. Exclusion of mesquite as a separate canopy type from the simplified habitat classification selected for mourning dove may account for the low multiple correlation coefficient associated with models for the South Texas Plains and Rolling Plains. Selection of mesquite habitats by nesting doves on the Rolling Plains has been reported (Jackson 1940). Inclusion of mesquite as a separate canopy type appears to be important in evaluating the habitat of nesting mourning doves within these ecological areas, as well as habitats of both species of quail throughout most of Texas during the breeding season. Similarly, absence of ground cover characteristics within the simplified habitat classification selected for bobwhite and scaled quail may have resulted in low multiple correlation coefficients associated with models for the Edwards Plateau, Rolling Plains, and High Plains, ecological areas in which overgrazing or cultivation was extensive. Ground cover height

and composition, however, appear not to be as important as the spatial distribution and composition of the canopy in evaluating habitats of the three species during the breeding season in Texas.

Results of multiple regression and discriminant analyses suggest that habitat variables might be used to predict audio counts of the three species and identify areas supporting above or below average densities within most of the ecological areas of Texas. Reasons for the low multiple correlation coefficients associated with models for mourning dove call counts within the Gulf Prairies and Marshes and scaled quail whistle counts within the Trans-Pecos are not known. High habitat heterogeneity in conjunction with low and uniform mourning dove call counts within the Gulf Prairies and Marshes (Grue 1977) may account for the low multiple correlation coefficient for this model. Conversely, habitat on transects within the Trans-Pecos was relatively homogeneous while whistle counts of scaled quail varied (Reid 1977), suggesting our habitat classification may not have included an important component of the habitat which this species selected for nesting.

Several improvements may be useful in testing and applying the techniques presented. Two improvements could be made in methods used to inventory habitat parameters. Habitat data could be collected using "mark-sense" computer data forms which would eliminate manual transfer of data to computer cards. Collection of habitat data in the sequence observed would permit direct calculation of habitat interspersion at any level within the habitat classification. In addition, aerial photographs and satelite imagery may facilitate inventory of habitat data, particularly within remote areas.

Data analysis could be improved by increasing the number of transects within physiographic units so that the number of observations exceeds the number of habitat types within the habitat classification selected. In the present study, we included audio-count data for all surveys conducted on the 133 call-count transects because differences in audio counts between surveys were significant. By including variation in call counts between surveys, high multiple correlation coefficients were more difficult to obtain (Grue 1977). However, inclusion of multiple surveys of individual transects artificially increased sample sizes within ecological areas, as surveys of the same transect were not statistically independent.

Limitations of stepwise multiple regression analyses should also be considered. Habitat parameters within models may not be the only ones significantly correlated with density. In addition, individual regression coefficients may depend on other variables within a model. Examination of correlation matrices may, therefore, prove useful in 1) identifying all habitat variables significantly correlated with density of the three species, 2) identifying habitat variables which may be more easily inventoried than those within the models and substituted without a significant reduction in multiple correlation coefficients, and 3) establishing guidelines for improving habitat for nesting mourning dove and bobwhite and scaled quail.

Use of the techniques presented to predict fluctuations in wildlife density and evaluate and manage wildlife habitat are outlined in figure 2. Procedures may be divided into three stages: 1) development of multiple linear regression models and discriminant functions which account for the greatest amount of variation in density indices and classes, respectively, and identification of habitat parameters significantly correlated with density of the wildlife species of interest; 2) testing of multiple linear regression models and discriminant functions; and 3) use of models to predict fluctuations in wildlife density, or use of discriminant functions and simple correlation analyses to evaluate and manage wildlife habitat. Procedures within the first stage have already been discussed here and elsewhere (Grue 1977, Reid 1977, Reid et al. 1978, 1979).

Multiple linear regression models may be tested by inventorying habitat parameters and the wildlife species of interest simultaneously on transects within the physiographic unit the models represent and comparing predicted densities and their confidence limits with observed values through time. If predicted and observed densities remain similar over time (i.e., habitat condition is the major factor governing fluctuations in density), the model(s) may be used to predict fluctuations in density. However, if predicted and observed densities differ significantly over time, other environmental factors (e.g., weather, disease, hunting) may be governing fluctuations in the density of species.

Discriminant functions should be tested by inventorying both habitat parameters and wildlife species of interest on additional transects within the physiographic unit for which each function was developed. If the percentage of the test transects correctly classified into each density class is high, the functions may be used to evaluate the habitat intersecting other transects within the physiographic unit for which they were developed. In areas for which the predicted density class of the wildlife species of interest is low, management recommendations could be made to increase those habitat parameters positively correlated with density of the species.

CONCLUSIONS

We believe the habitat classification and techniques described may be applicable to evaluating habitats of breeding mourning dove and bobwhite and scaled quail throughout their ranges, and habitats of other wildlife species for which line transects are used to collect population data. Methods presented may prove useful in

predicting annual fluctuations in wildlife density, in determining effects of habitat modification on wildlife density, and in evaluating and managing wildlife habitat. Further research into development and testing of a windshield approach to the evaluation of wildlife habitat is needed and appears justified.

ACKNOWLEDGEMENTS

This study was funded by the U.S. Fish and Wildlife Service, the Caesar Kleberg Research Program in Wildlife Ecology, and The Agricultural Experiment Station, Texas A&M University, in cooperation with the Texas Parks and Wildlife Department. We gratefully acknowledge the assistance of F.W. Martin, Director, Migratory Bird and Habitat Research Laboratory, U.S. Fish and Wildlife Service, J.H. Dunks and J.T. Robertson, Texas Parks and Wildlife Department, and personnel of the Texas Parks and Wildlife Department who conducted the audio counts. We are also indebted to J.L. Folse and W.E. Grant for review of the manuscript. This paper constitutes part of a dissertation and master's thesis by the senior and second authors, respectively.

LITERATURE CITED

Amend, S.R. 1969. Progress report on Carolina Sandhills mourning dove studies. Proceedings Annual Conference Southeastern Association Game and Fish Commissioners 23:191-201.

Barr, A.J., and J.H. Goodnight. 1972. A user's guide to the statistical analysis system. 260 p. SAS Institute, Inc. Raleigh, N. Car.

Baxter, W.L., and C.W. Wolfe. 1973. The interspersion index as a technique for evaluation of bobwhite quail habitat. p. 158-165. In Morrison, J.A. and J.C. Lewis, editors. Proceedings First National Bobwhite Quail Symposium [Stillwater, Okla., April 23-26, 1972]. Oklahoma State University.

Bennitt, R. 1951. Some aspects of Missouri quail and quail hunting, 1938-1948. 51 p. Missouri Conservation Commission Technical Bulletin 2.

Blankenship, L.H., A.B. Humphrey, and D. MacDonald. 1971. A new stratification of mourning dove call-count routes. Journal of Wildlife Management 35:319-326.

Brown, D.E., C.L. Cochran, and T.E. Waddell. 1978. Using call-counts to predict hunting success for scaled quail. Journal of Wildlife Management 42:281-287.

Campbell, H., D.K. Martin, P.E. Ferkovich, and B.K. Harris. 1973. Effects of hunting and some other environmental factors on scaled quail in New Mexico. 49 p. Wildlife Monograph 34.

Davey, P. 1953. The mourning dove in southern Michigan. M.S. Thesis. 47 p. University of Michigan, Ann Arbor, Mich.

Dolton, D.D. 1977. Mourning dove status report, 1976. 27 p. U.S. Fish and Wildlife Service Special Scientific Report Wildlife 208.

Dunks, J.H. 1976. Statewide mourning dove research; Job No. 1: Density, distribution, and movement; Texas. Report on Federal Aid Project W-95-R-10. 9 p. Texas Parks and Wildlife Department.

Elder, J.B. 1956. Analysis of whistling patterns in the eastern bobwhite, Colinus v. virginianus L. Proceedings Iowa Academy Science 63:639-651.

Foote, L.E., H.S. Peters, and A.L. Finkler. 1958. Design tests for mourning dove call count sampling in seven southeastern states. Journal of Wildlife Management 22:402-408.

Gould, F.W. 1975. Texas plants - a checklist and ecological summary. 121 p. Texas A&M University Agricultural Experiment Station. MP-585/Rev.

Grue, C.E. 1977. Classification, inventory, analysis, and evaluation of the breeding habitat of the mourning dove (Zenaida macroura) in Texas. Ph.D. Thesis. 187 p. Texas A&M University, College Station, Tex.

Grue, C.E., R.R. Reid, and N.J. Silvy. 1976. A technique for evaluating the breeding habitat of mourning doves using call-count transects. Proceedings Annual Conference Southeastern Association Game and Fish Commissioners. 30:667-673.

Jackson, A.S. 1940. The mourning dove in Throckmorton County, Texas. M.S. Thesis. 122 p. North Texas State University, Denton, Tex.

Keeler, J.E., chairman. 1977. Mourning dove (Zenaida macroura). p. 275-298. In Sanderson, G.C., editor. Management of migratory shore and upland game birds in North America. 358 p. International Association Fish and Wildlife Agencies, Washington, DC.

Norton, H.W., T.G. Scott, W.R. Hanson, and D.W. Klimstra. 1961. Whistling-cock indices and bobwhite populations in autumn. Journal of Wildlife Management 25:398-403.

Reid, R.R. 1977. Correlation of habitat parameters with whistle-count densities of bobwhite (Colinus virginianus) and scaled quail (Callipepla squamata) in Texas. M.S. Thesis. 101 p. Texas A&M University, College Station, Tex.

Reid, R.R., C.E. Grue, and N.J. Silvy. 1978. Breeding habitat of the bobwhite in Texas. Proceedings Annual Conference Southeastern Fish and Wildlife Agencies 32:62-71.

Reid, R.R., C.E. Grue, and N.J. Silvy. 1979. Competition between bobwhite and scaled quail in Texas. Proceedings Annual Conference Southeastern Association Fish and Wildlife Agencies 33:146-153.

Rosene, W., Jr. 1957. A summer whistling cock count of bobwhite quail as an index to wintering populations. Journal Wildlife Management 21:153-158.

Sandfort, W.W. 1977. Introduction. p. 1-3. In Sanderson, G.C., editor. Management of migratory shore and upland game birds in North America. 358 p. International Association Fish and Wildlife Agencies, Washington, D.C.

Shannon, C.E. 1948. A mathematical theory of communication. Bell System Technical Journal 27:379-423; 623-656.

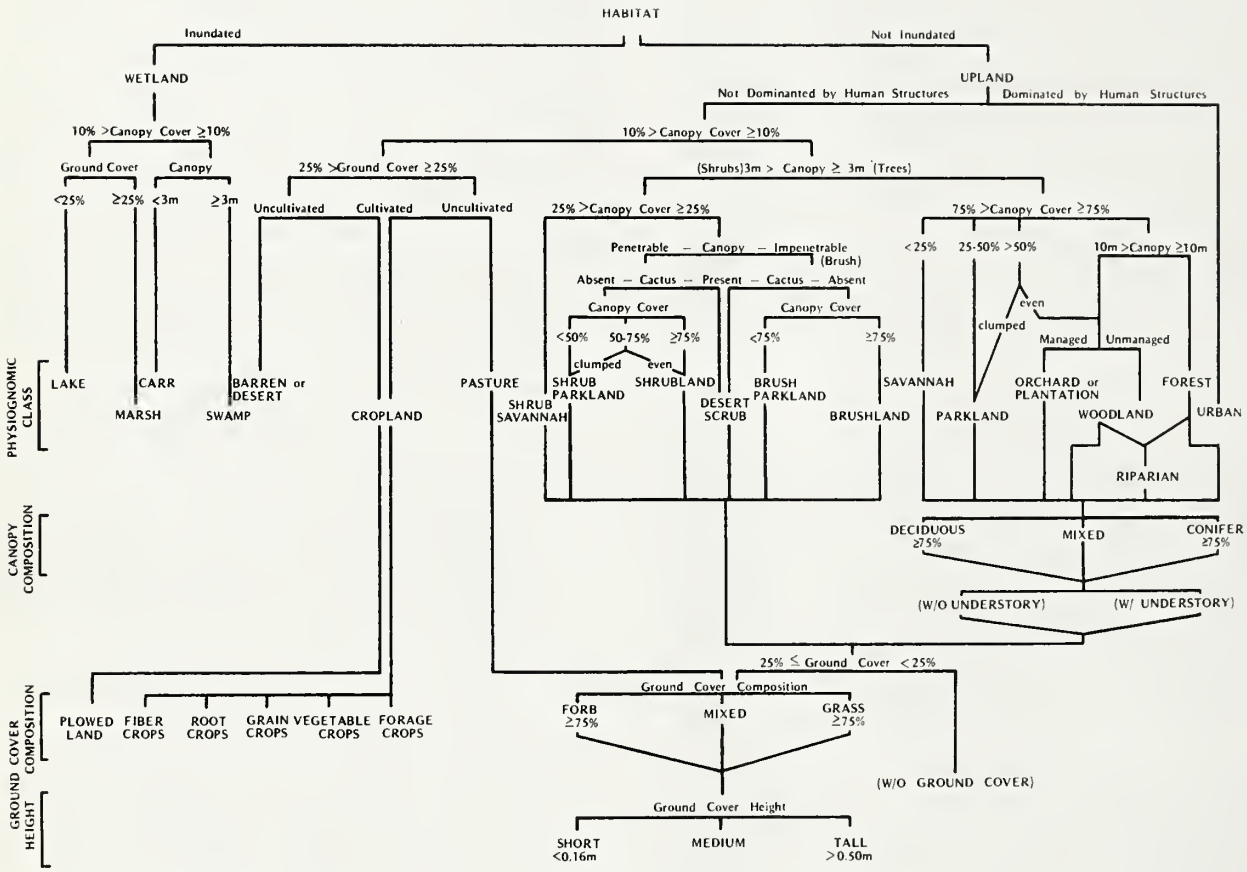


Figure 1. Schematic diagram of habitat classification (after Grue et al. 1976). Habitats were keyed to habitat types from top to bottom. Names were assigned habitat types from bottom to top using words in bold capital letters describing the height and compostion of the ground cover, composition of the canopy, and the physiognomic class, in that order (e.g., tall grass deciduous savannah). Words in brackets specifying the presence or absence of ground cover or understory comprised the last portion of the name for habitat types where appropriate (e.g., mixed woodland with understory).

Table 1. Mean transect audio counts for mourning dove (MD), bobwhite quail (BW), and scaled quail (SQ) within the 10 ecological areas of Texas in 1977. Means are rounded to nearest whole bird. N represents the number of valid surveys conducted.

	MD			BW			SQ		
Ecological area	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD
Pineywoods	32	10	7.4	23	13	12.8	23	0	--
Gulf Prairies and Marshes	18	7	3.8	13	43	16.4	13	0	--
Post Oak Savannah	33	17	13.2	24	30	19.5	24	0	--
Blackland Prairies	40	17	9.5	28	29	13.2	28	0	--
Cross Timbers and Prairies	62	27	21.3	44	46	27.7	44	0	--
South Texas Plains	70	28	17.6	52	27	18.9	52	2	3.8
Edwards Plateau	72	18	12.7	54	12	15.6	54	5	8.3
Rolling Plains	88	45	26.3	64	38	20.4	64	4	7.1
High Plains	52	7	7.9	38	6	8.2	38	3	4.7
Trans-Pecos	35	16	18.9	26	0		26	10	5.0

Table 2. Correlations between audio counts of mourning dove (MD), bobwhite quail (BW), and scaled quail (SQ) and habitat variables including habitat types within the initial habitat classification and structural features. Multiple correlation coefficients are expressed as a percent. The number of variables remaining in the models is given in parentheses.

Ecological area	No. habitat variables present	MD	BW	SQ
Pineywoods	98	83.3 (3)	91.1 (3)	
Gulf Prairies and Marshes	96	28.3 (1)	82.2 (2)	
Post Oak Savannah	126	85.9 (3)	94.3 (4)	
Blackland Prairies	105	68.1 (3)	58.9 (3)	
Cross Timbers and Prairies	145	79.3 (6)	83.6 (5)	
South Texas Plains	117	78.2 (4)	93.9 (8)	92.6 (4)
Edwards Plateau	66	64.1 (6)	85.7 (6)	92.9 (5)
Rolling Plains	92	73.5 (10)	89.2 (10)	78.1 (4)
High Plains	52	62.9 (2)	93.8 (5)	81.7 (4)
Trans-Pecos	36	86.5 (3)		31.9 (1)
Texas	194	79.8 (54)	89.5 (51)	65.8 (16)

Table 3. Correlations between transect audio counts of mourning dove (MD), bobwhite (BW), and scaled quail (SQ) and habitat variables considering structural features and habitat types within the initial habitat classification, separately. Multiple correlation coefficients are expressed as percents. The number of variables remaining in the models is given in parentheses.

	Structural features				Habitat types			
Ecological area	Number present	MD	BW	SQ	Number present	MD	BW	SQ
Pineywoods	19	80.2(3)	91.1(5)		75	83.2(3)	91.1(3)	
Gulf Prairies and Marshes	19	28.3(1)	80.0(2)		73	22.8(1)	82.2(2)	
Post Oak Savannah	20	87.0(4)	80.7(2)		102	85.9(3)	94.9(5)	
Blackland Prairies	20	47.5(1)	46.2(2)		81	70.8(4)	58.9(3)	
Cross Timbers and Prairies	20	78.4(10)	65.5(5)		121	79.3(6)	83.5(6)	
South Texas Plains	22	70.2(8)	76.1(6)	92.6(3)	91	78.2(4)	94.3(9)	92.7(5)
Edwards Plateau	16	32.7(4)	54.7(6)	44.8(3)	47	64.3(6)	83.6(5)	93.2(10)
Rolling Plains	19	47.7(5)	40.9(5)	44.8(3)	70	72.1(10)	86.5(10)	78.0(4)
High Plains	17	51.3(4)	75.2(4)	74.7(3)	32	62.9(2)	93.5(5)	81.7(4)
Trans-Pecos	16	84.1(3)		31.9(1)	17	81.5(2)		30.0(1)
Mean	19	60.7(4)	67.8(4)	57.8(3)	71	70.1(4)	85.4(5)	75.1(5)

Table 4. Correlations between transect audio counts of mourning dove, bobwhite quail, and scaled quail and habitat types within simplified habitat classifications. Multiple correlation coefficients are expressed as percents; those underlined represent the habitat classification scheme selected as 'best.' The number of potential habitat types within each classification scheme is given in parentheses.

			Physiognomic class						
					w/canopy composition			w/understory	
Species	Ecological area	Initial (501)	w/mesquite (19)		(33)	(29)	w/crops (34)	w/mesquite (43)	(44)	w/crops (49)
Mourning Dove	Pineywoods	83.2	83.3		83.3	82.9	82.9	82.9	82.2	82.2
	Gulf Prairies & Marshes	22.8	0.0		0.0	0.0	22.8	0.0	0.0	22.8
	Post Oak Savannah	85.9	86.7		85.8	87.6	85.9	84.5	85.5	85.9
	Blackland Prairies	70.8	70.0		71.0	70.2	70.2	65.2	69.3	56.2

(continued)

(table 4 continued)

	Cross Timbers & Prairies	79.3	61.9	68.9	71.7	78.4	79.1	73.1	79.0
	South Texas Plains	78.2	35.0	74.3	28.2	60.5	75.6	29.3	60.4
	Edwards Plateau	64.3	42.6	42.6	55.1	55.1	57.8	55.1	55.1
	Rolling Plains	72.1	35.8	40.1	36.4	36.4	50.2	36.4	36.4
	High Plains	62.9	64.0	64.0	64.0	63.9	64.0	64.0	64.0
	Trans-Pecos	81.5	80.7	24.3	80.7	81.5	24.3	80.7	81.5
	Mean	70.1	56.0	55.4	57.6	63.8	58.4	57.6	62.3
Bobwhite Quail	Pineywoods	91.1	90.6	90.6	88.7	88.7	88.7	90.2	90.2
	Gulf Prairies & Marshes	82.2	79.3	81.9	80.5	69.3	81.9	80.4	69.3
	Post Oak Savannah	94.9	92.2	87.9	94.2	94.4	89.0	94.2	88.4
	Blackland Prairies	58.9	0.0	63.7	59.4	16.6	64.7	52.9	38.5
	Cross Timbers & Prairies	83.5	0.0	58.7	0.0	60.2	79.1	0.0	60.2
	South Texas Plains	94.4	68.8	69.4	68.1	79.8	94.7	68.1	79.8
	Edwards Plateau	83.6	75.3	79.7	75.6	75.6	81.5	75.6	75.6
	Rolling Plains	86.5	46.4	72.5	50.2	57.1	70.3	52.8	52.8
	High Plains	93.5	75.5	75.4	75.5	80.0	72.4	88.6	88.6
	Mean	85.4	58.7	75.5	65.8	69.1	80.3	67.0	71.5
Scaled Quail	South Texas Plains	92.7	19.5	70.1	19.5	19.5	85.3	19.5	19.5
	Edwards Plateau	93.2	44.6	36.9	44.6	44.6	42.0	44.6	44.6
	Rolling Plains	78.0	58.7	67.3	58.7	58.7	67.3	58.7	58.7
	High Plains	81.7	66.2	81.7	66.2	74.6	81.7	70.3	74.4
	Trans-Pecos	30.0	0.0	30.0	0.0	15.5	30.0	0.0	15.5
	Mean	75.1	37.8	57.2	37.8	42.5	61.3	38.6	42.5

Table 5. Multiple linear regression models for transect audio counts of mourning dove (MD), bobwhite quail (BW), and scaled quail (SQ) and habitat variables including habitat types within the simplified habitat classifications selected and structural features. Multiple correlation coefficients are expressed as percents.

Ecological Area	Model	R^2
Pineywoods	MD = -0.4 + 0.8[PARALLEL SHRUBROWS] - 4.0[DECIDUOUS FOREST] + 0.4[PARALLEL POWERLINES]	81.3
	BW = -15.7 + 12.4[DECIDUOUS SAVANNAH] + 9.5[DECIDUOUS FOREST] +0.3[PARALLEL FENCES]	90.9
Gulf Prairies & Marshes	MD = 1.3 + 0.1[INTERSECTING FENCES]	28.3
	BW = 22.5 + 41.5[SAND ROAD SURFACE] + 18.2[SHRUB SAVANNAH]	82.0
Post Oak Savannah	MD = 28.0 + 18.8[HAY] - 7.0[GRAVEL PITS] + 10.3[SHRUB PARKLAND] - 1.4[PASTURE OR FIELDS]	87.5
	BW = 11.6 + 54.5[MESQUITE WOODLAND] + 33.9[MIXED MESQUITE SHRUB PARKLAND] + 101.6[CONIFER PARKLAND] + 307.4[ORCHARDS]	93.8
Blackland Prairies	MD = 55.6 - 1.1[INTERSECTING ROADS] + 86.7[ORCHARDS] -0.1[BUILDINGS AND ASSOCIATED VEGETATION]	68.1
	BW = 19.6 - 99.1[MESQUITE SHRUB SAVANNAH] + 0.5[BUILDINGS AND ASSOCIATED VEGETATION] + 1.1[PARALLEL POWERLINES]	60.1
Cross Timbers & Prairies	MD = 98.8 + 6.1[GRAVEL PITS] - 0.3[INTERSECTING FENCES] + 6.8[DECIDUOUS WOODLAND] - 5.0[WASHES] + 108.4[CONIFER PARKLAND] - 37.0[HABITAT DIVERSITY] + 3.4[PLOWED LAND]	71.5
	BW = 611.8 + 22.5[PARALLEL FENCES] - 181.7[BARREN LAND] - 59.3[SHRUB PARKLAND] + 0.8[ROAD SHOULDER WIDTH] + 5.8 [GRAVEL PITS] - 259.2[MIXED MESQUITE SHRUBLAND] + 34.3[MIXED MESQUITE SHRUB SAVANNAH]	77.9
South Texas Plains	MD = 7.6 + 19.5[HAY] - 4.9[PASTURE OR FIELDS] + 2.2[SNAGS] - 0.6[INTERSECTING SHRUBROWS] + 14.8[HABITAT DIVERSITY] - 2.1[DECIDUOUS PARKLAND]	74.7
	BW = 12.0 + 5.1[MESQUITE PARKLAND] + 12.0[HABITAT DIVERSITY] - 2.0[GRAVEL ROAD SURFACE] + 3.0[LIVESTOCK FEEDERS] + 5.0[BRUSHLAND] + 5.9[INTERSECTING RAILROADS] + 0.4[BRUSH W/MESQUITE] + 5.7[DECIDUOUS WOODLAND] + 1.0[MIXED MESQUITE SHRUBLAND] + 0.4[PARALLEL POWERLINES]	95.0
	SQ = 8.9 - 0.2[INTERSECTING POWERLINES] + 0.3[PARALLEL WINDBREAKS] + 4.1[URBAN DEVELOPMENT] - 1.8[INTERSECTING RAILROADS] - 2.9[IRRIGATION PUMPS] + 2.6 [BRUSHLAND] - 0.2[ROAD SHOULDER WIDTH] - 1.5[DECIDUOUS SAVANNAH] - 0.3 [MESQUITE PARKLAND] + 0.2[ASPHALT ROAD SURFACE] + 0.3[SHRUBLAND]	93.1
Edwards Plateau	MD = 2.9 + 14.5[URBAN DEVELOPMENT] + 4.3[BRUSHLAND] + 4.9[MIXED PARKLAND] + 2.9 [PASTURE OR FIELDS] + 0.3[INTERSECTING FENCES] - 1.1[MIXED WOODLAND] + 5.2 [INTERSECTING RAILROADS]	63.9
	BW = 1.9 + 7.6[BRUSHLAND] + 3.8[DECIDUOUS SAVANNAH] + 2.2[MESQUITE WOODLAND] - 1.9[MESQUITE SHRUBLAND]	81.5
	SQ = 7.2 - 3.9[HABITAT DIVERSITY] + 6.5[SHRUB SAVANNAH] + 17.3[INTERSECTING RAILROADS] + 0.4[SHRUBLAND] - 1.0[WASHES] - 0.7[INTERSECTING ROADS] + 0.4 [DECIDUOUS SAVANNAH] + 0.1[PARALLEL POWERLINES]	93.3

(continued)

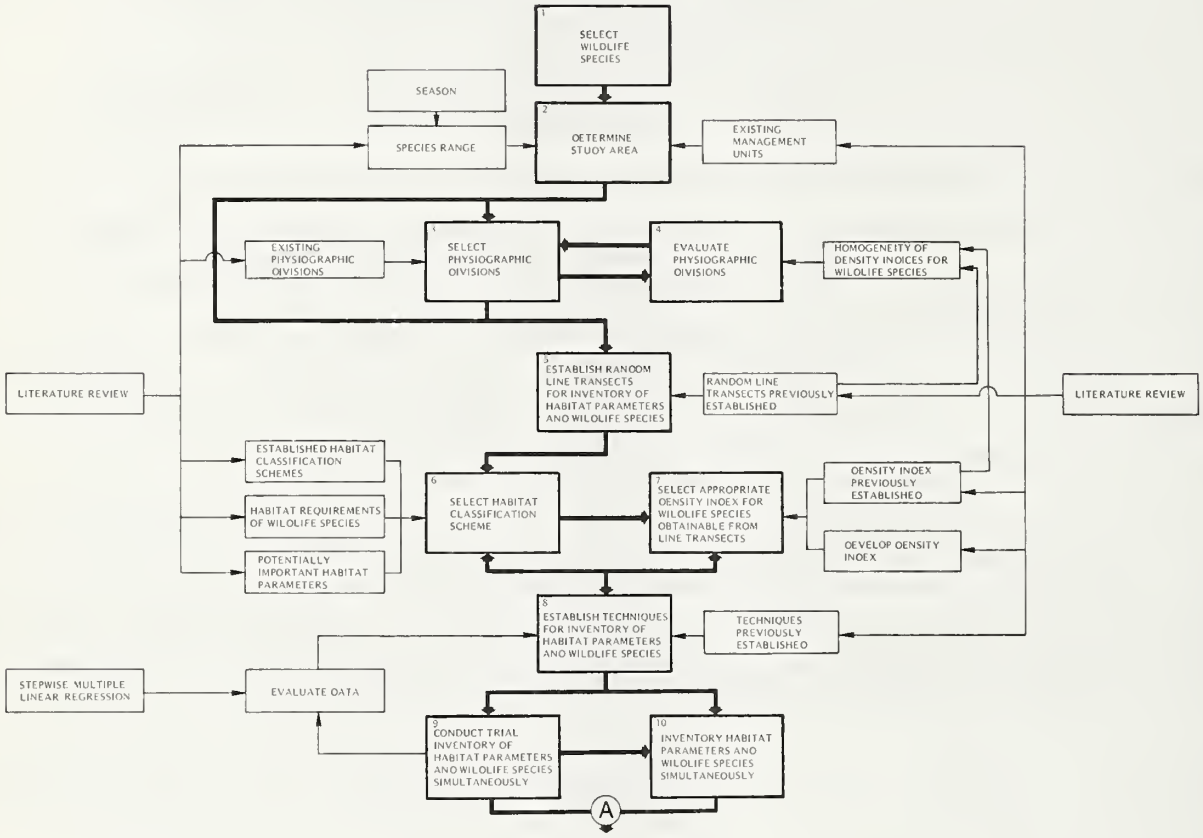
(table 5 continued)

Rolling Plains	MD = 2.1 + 3.4[DECIDUOUS PARKLAND] + 2.0[SNAGS] + 26.0[HABITAT DIVERSITY] - 3.3 [PARALLEL WINDBREAKS]	55.7
	BW = - 18.6 + 66.8[HABITAT DIVERSITY] + 33.1[MESQUITE SHRUB PARKLAND] + 4.1 [INTERSECTING WINDBREAKS] - 1.7[INTERSECTING ROADS] - 2.3[MESQUITE SAVANNAH] + 98.3[CONIFER SAVANNAH] + 3.4[BUILDINGS AND ASSOCIATED VEGETATION] - 2.7 [INTERSECTING POWERLINES] - 371.1[ORCHARDS] + 198.4[CONIFER PARKLAND] +3.1[MESQUITE PARKLAND] - 5.9[MIXED MESQUITE PARKLAND] + 56.2[DECIDUOUS WOODLAND] - 1.3[HABITAT INTERSPERSION]	87.4
	SQ = - 4.5 + 2.1[WASHES] + 4.2[SHRUB PARKLAND] - 11.3[MESQUITE SHRUBLAND] + 0.4 [PARALLEL POWERLINES] + 3.8[MESQUITE SHRUB PARKLAND] + 1.7[INTERSECTING RAILROADS]	75.4
High Plains	MD = 0.3 + 71.9[DECIDUOUS SAVANNAH] + 0.7[PASTURE OR FIELDS]	63.9
	BW = 3.0 + 62.8[SHRUBLAND] + 0.9[GRAVEL ROAD SURFACE] - 32.0[SHRUB PARKLAND] + 0.3[BUILDINGS AND ASSOCIATED VEGETATION] - 0.6[ROAD SHOULDER WIDTH] - 3.3 [MESQUITE SHRUB SAVANNAH]	92.6
	SQ = 0.7 + 25.7[MESQUITE SHRUBLAND] + 74.8[SHRUB SAVANNAH] + 33.2[SHRUB PARKLAND] + 4.7[URBAN DEVELOPMENT]	81.7
Trans-Pecos	MD = 16.1 + 23.5[PLOWED LAND] - 4.7[BUILDINGS AND ASSOCIATED VEGETATION] - 0.5 [HABITAT INTERSPERSION]	86.6
	SQ = 8.9 + 0.6[IRRIGATION AND OIL PUMPS]	31.9

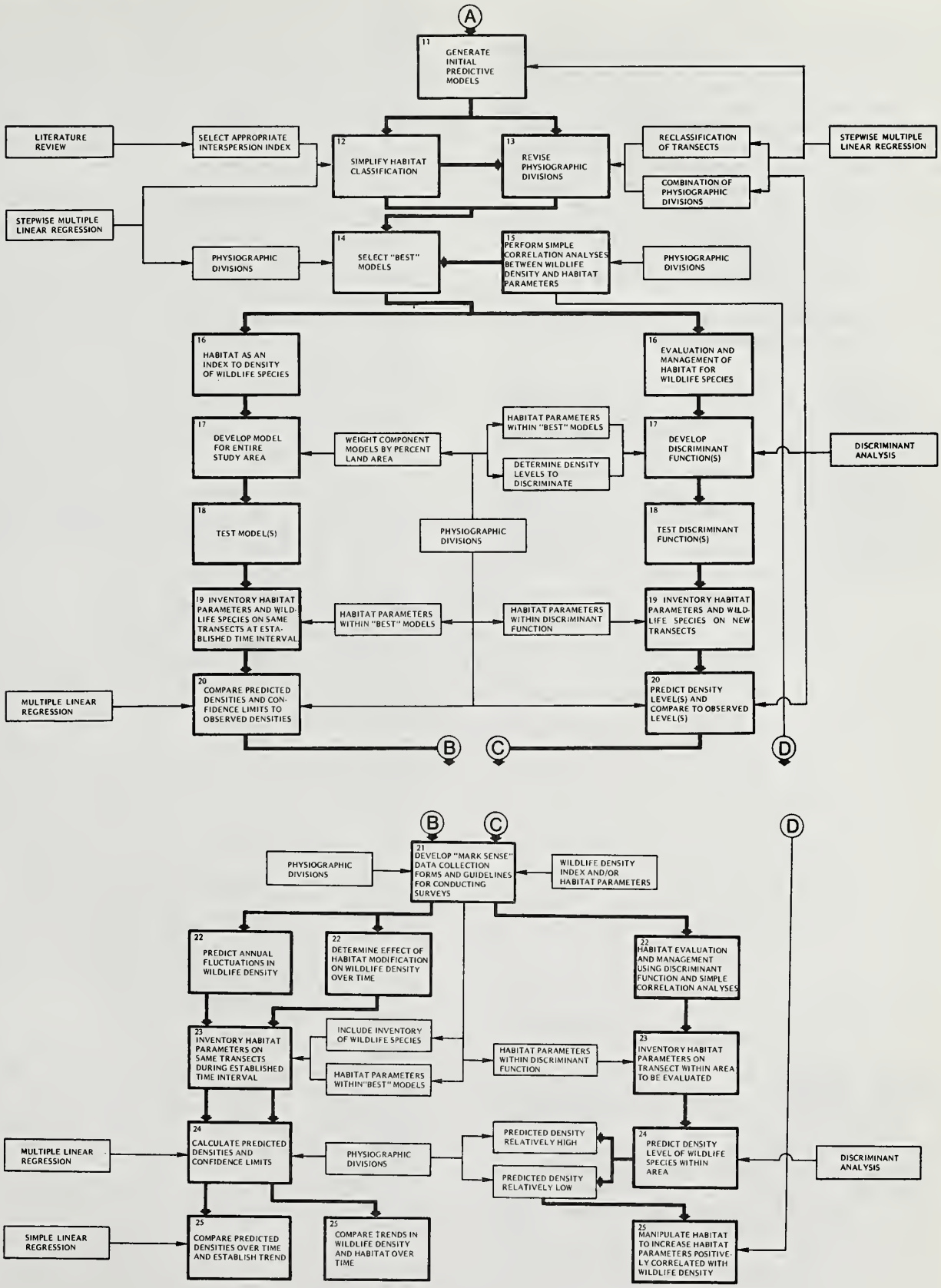
Table 6. Percent audio-count surveys of mourning dove (MD), bobwhite quail (BW) and scaled quail (SQ) correctly classified as below average or average and above by discriminant analyses. Discriminant functions incorporated the habitat variables within multiple linear regression models for each ecological area (see table 5); differences in habitat variables between audio-count classes were significant (MANOVA: $P<0.05$). N represents the number of audio count surveys.

Ecological area	MD		BW		SQ
	N	%	N	%	%
Pineywoods	32	81.3	23	95.7	
Gulf Prairies and Marshes	18	77.8	13	84.6	
Post Oak Savannah	33	84.8	24	70.8	
Blackland Prairies	40	82.5	28	82.1	
Cross Timbers and Prairies	62	85.5	44	90.9	
South Texas Plains	70	84.3	52	94.2	96.2
Edwards Plateau	72	77.8	54	74.1	98.1
Rolling Plains	88	79.5	64	95.3	93.8
High Plains	52	75.0	38	89.5	89.5
Trans-Pecos	35	88.6	26		38.5
Mean		81.7		86.4	83.2

Figure 2. Flow diagram for the development, testing, and use of a windshield approach to the evaluation of wildlife habitat.



(figure 2 continued)



DISCUSSION

PAUL GEISSLER: I have two comments. First, you have an average of about eight times as many variables as transects in your initial analysis. Even counting the artificially increased sample sizes which resulted from using multiple surveys of transects as independent observations, you have an average of over twice as many variables as observations. In addition to the variables examined in your initial analysis, you also examined additional variables that were combinations of the original variables in the process of evaluating seven simplifications of habitat classifications. It does not matter that simplified habitat classifications were evaluated in separate computer runs, because these additional variables were still included in the search for a simple model with high predictive power.

Very high R^2's can be obtained from random numbers which have no predictive power whatsoever (figure 1D). This is due essentially to the repeated analyses on the same set of data by the stepwise procedure. When there are few observations relative to the number of variables, there is a tendency for the procedure to fit the random variations in the data as well as the underlying biological process. This results in prediction bias, the over-estimation of the

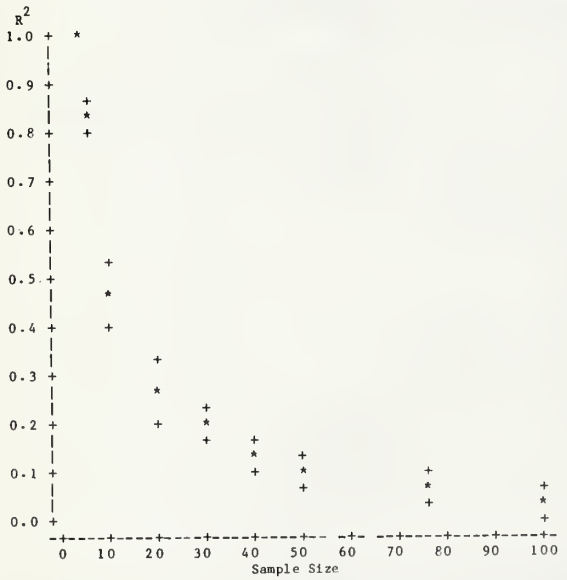


Figure 1D. Effect of sample size on R^2 in stepwise regression. One hundred sets of data were generated for each of the sample sizes of 3, 5, 10, 20, 30, 40, 50, 75, and 100 observations. Each observation consisted of a dependent variable and 10 independent variables generated as normally distributed random numbers. Each set of data was analyzed using the SAS Stepwise Procedure with default options (Barr et al., op. cit.) and the mean R^2 and the 95% confidence limits for the mean R^2 plotted as a function of sample size.

model's predictive ability (R^2) based on the data used to construct the model as compared to the model's true predictive ability on other data (Neter, J., and W. Wasserman. 1974. Applied linear statistical models. Richard D. Irwin, Inc., Homewood, Ill. p.388). In other words, a very high R^2 indicating good predictive power may be obtained from the original set of data, but poor predictions and low R^2 result when the model is applied to another set of data.

Using the Pineywoods as an example, you obtained a R^2 of 83.3% using a model with three variables selected from 98 variables present. There were nine transects in this area. Using independent normally distributed (0,1) random numbers in place in these 98 prediction variables and the response variable for nine independent transects, I calculated an R^2 of 100% 24 out of 25 times using the SAS Stepwise Procedure with default options (Barr, A.J. et al. 1979. SAS User's Guide. SAS Institute, Cary, N. C.). There are similar problems with the discriminant analysis unless there are substantially more independent observations than there are variables.

My second comment concerns the correlation among your call-count surveys. You indicate that audio count data for all of the surveys conducted on the 133 call-count transects were included instead of using transect means. This was done because differences in audio counts between surveys were significant. However, you note that "...inclusion of multiple surveys of individual transects artificially increased sample sizes within ecological areas, as surveys of the same transect were not statistically independent." A basic assumption in any regression analysis is that the error terms are uncorrelated (Neter and Wasserman, op. cit., p. 31). The test for a difference among surveys is not relevant to the determinations of the unit that constitutes an independent observation. That decision must be based on the lack of correlation among units. It is possible for there to be a perfect correlation (r=1) between the call-counts on multiple surveys of individual transects, and at the same time be significant differences among the surveys. What you need to demonstrate is that the survey results are uncorrelated, not that there is a significant difference among surveys. Analyzing the mean of the multiple surveys of a transect will result in uncorrelated observations and meet the assumptions required for regression analysis.

CHRIS GRUE: We were aware of the possibility of prediction bias throughout our study. The number of potential habitat types within our habitat classification was, unfortunately, nearly four times the number of call-count transects within Texas. We, however, considered the habitat classification to be suitable for classifying habitats from within a vehicle and an unbiased method of selecting the habitat variables to be included in the regression analyses; the classification scheme was developed prior to the collection of any field data. We believed the elimination of levels within the classification

hierarchy was an objective means of simplifying the habitat classification and reducing the number of habitat variables therein. We also realized that, depending on the level within the habitat classification, only a fraction of the potential habitat variables might actually be observed along the call-count transects. With respect to the discriminant analyses, we decided to include only those habitat variables which entered the regression models.

Because the number of call-count transects was lower than the potential number of habitat variables, the decision of whether to conduct the regression analyses statewide or within ecological areas was difficult. Both approaches had apparent problems, statistical or biological. By conducting regression analyses statewide, the potential for prediction bias would be reduced. However, this approach appeared to suffer biologically. Habitat variables important to doves may not be expected to be the same within any two ecological areas due to differences in the abundance and distribution of habitat types. Analyses within ecological areas, though probably more valid biologically, increased the potential for prediction bias; we were restricted to sampling the existing call-count transects due to budget constraints. Since the two approaches appeared to have merit, we decided, apriori, to utilize both in the study.

To counter the possible effects of prediction bias in the regression analyses conducted within ecological areas, we included call-counts for all surveys as the dependent variable instead of transect means. Variability in call-counts between surveys was great and we believed its inclusion in the regression analyses would make it more difficult to obtain high multiple correlation coefficients. Because surveys of the same transect were not independent observations and their inclusion artificially increased sample size, we increased the significance level for entry of habitat variables into the models from the default value of $P<0.10$ (Barr et al., op. cit.) to $P<0.05$.

Multiple correlation coefficients for models from statewide and within-ecological-area analyses using call-counts for all surveys and transect means are presented in table 1D. Inclusion of call-counts for all surveys did make it more difficult to obtain high multiple correlation coefficients. Some prediction bias is undoubtedly present in the data, particularly models for ecological areas based on transect means. The amount of prediction bias present, however, cannot be assessed until the models are tested. We believe the amount of prediction bias present is

Table 1D. Correlations between mourning dove call-counts (all surveys conducted and transect means) and habitat variables (structural features and habitat types within the simplified habitat classification selected). Multiple correlation coefficients are expressed as percents. The number of habitat variables remaining in the models is given in parentheses.

Ecological area	No. habitat variables present	All surveys N	All surveys R^2	Mean of surveys N	Mean of surveys R^2
Pineywoods	54	32	81.3 (3)	9	100.0 (6)
Gulf Prairies and Marshes	40	18	28.3 (1)	6	100.0 (3)
Post Oak Savannah	49	33	87.5 (4)	9	100.0 (6)
Blackland Prairies	44	40	68.1 (3)	10	93.6 (3)
Cross Timbers and Prairies	46	62	71.5 (7)	17	27.0 (1)
South Texas Plains	43	70	74.7 (6)	18	86.1 (4)
Edwards Plateau	35	72	63.9 (7)	18	81.7 (4)
Rolling Plains	37	88	55.7 (4)	23	56.6 (2)
High Plains	34	52	63.9 (2)	14	89.5 (2)
Trans-Pecos	31	35	86.6 (3)	9	99.7 (4)
Texas	54	502	44.6(16)	133	38.2 (4)

substantially less than that suggested by your example. One would not expect to account for only 27% of the variation in call-counts given 17 transects and 46 habitat variables if prediction bias was severe. If one is willing to accept the results of our regression analyses in which the number of independent observations substantially exceeded the number of independent variables (e.g., 3X), then habitat variables still accounted for a significant portion of the variation in mourning dove call counts (ca. 40%) throughout Texas (table 1D). That regression models for ecological areas accounted for a greater proportion of the variation in call counts than the model for the State, however, appears to be reasonable biologically. Prediction bias in our discriminant analyses should not be great; the number of transects was two to seven times the number of habitat variables included.

BERNARD MORZUCH: What are the implications of using stepwise regression on the properties of the resulting parameter estimates in the regression analyses?

CHRIS GRUE: Compared to "all possible regression" procedures, predictive models generated by stepwise may not provide the best fit to a data set because a limited amount of stepping back is done. Models generated by either procedure may not include all independent variables significantly correlated with a dependent variable and individual regression coefficients may depend on the other parameters within a model. Models generated by either procedure are primarily restricted to predictive uses. Examination of correlation matrices may, therefore, prove useful in identifying all habitat parameters significantly correlated with wildlife density and in establishing guidelines for improving wildlife habitat.

Examples: Multivariate Analyses of Wildlife Habitats

INTERSPECIFIC DIFFERENCES IN NESTING HABITAT OF SYMPATRIC WOODPECKERS AND NUTHATCHES[1]

Martin G. Raphael[2]

Abstract.--To test for nest site differences among nine sympatric species of woodpeckers and nuthatches (hole excavators) in a Sierra Nevada mixed conifer forest, I located 306 active nests, measured forest stand characteristics on a 0.04 ha plot centered at each nest, measured characteristics of the nest tree, and used discriminant analysis to compare these nest site characteristics among bird species.

Three discriminant functions were considered. The first was associated most strongly with live tree basal area and canopy height. The second function was associated with nest tree species and nest tree height; the third was identified by nest tree diameter and top condition. Mean discriminant scores differed significantly among all but 3 of 36 possible pairs of species along at least one of these discriminant axes, indicating that nearly all bird species chose distinct nest sites. The distributions of discriminant scores along each axis, however, showed considerable overlap with nest sites of two sapsucker species being the most similar. Euclidian distance between mean scores was the most useful measure to characterize the similarity of species' nest sites.

Nest stand and nest tree variables contributed nearly equally to the discrimination between bird species. This analysis suggested that both of these sets of variables should be included in management prescriptions to provide habitat for cavity nesting birds.

Key words: Cavity nesting birds; cluster analysis; discriminant analysis; Euclidian distance; nuthatches; Sierra Nevada; woodpeckers.

INTRODUCTION

Cavity nesting birds comprise about 30% of the breeding bird species in western forests. Habitat management for these species requires, among other factors, the provision of adequate numbers of suitable nest sites, usually standing dead trees (Raphael and White 1978, Thomas et al. 1979). The question arises whether each species selects distinct nest sites or whether groups of

[1]Paper presented at The use of multivariate statistics in the studies of wildlife habitat: a workshop, April 23-25 1980, Burlington, Vt.

[2]Staff Research Associate, Department of Forestry and Resource Management, University of California, Berkeley, CA 94720.

species nest in trees with similar characteristics. Habitat management is simplified in the latter case since species with similar nest sites can be grouped into a reduced set of functional management units.

The purpose of this study was to determine whether the nest sites of each species in a group of sympatric primary cavity nesters were distinct or whether one can identify subsets of species using nest sites with similar characteristics. To do so, I performed a one-way multivariate analysis of variance using discriminant analysis, followed by an examination of the relative separation of all possible species-pairs. The basic advantages of using discriminant analysis instead of a series of single-variable comparisons among all species are that it: a) accounts for correlations among the variables used in the analysis, and b) allows more rigorous control over the experiment-wise (type I) error rate.

Primary cavity nesters excavate their own cavities and to do so they must choose the appropriate substrate for the nest. Secondary cavity nesters, on the other hand, choose appropriate cavities, usually abandoned woodpecker holes (Raphael 1980). I limited the analysis to primary cavity nesters because these species select trees; the secondary cavity nesters choose cavities and this confounds an analysis of tree characteristics.

METHODS

Study Area

Field studies were conducted at the University of California Sagehen Creek Field Station, located on the east side of the Sierra Nevada, 13 km north and 6 km west of Truckee, California. Elevations in the 39 km^2 Sagehen Creek drainage vary from 1800 m to 2300 m. The drainage is dominated by a mix of Jeffrey pine (Pinus jeffreyi) and white fir (Abies concolor) and by brushfields or conifer plantations on the site of the 1960 Donner Ridge fire which burned the eastern quarter of the basin. Meadows, lodgepole pine (Pinus murryana), and aspen (Populus tremuloides) occur in mesic sites, and red fir (Abies magnifica) and mountain hemlock (Tsuga mertensiana) dominate at higher elevations.

Searches for active nests were conducted throughout the Sagehen Creek basin from 1976 to 1979. Active nests were confirmed by observing adults entering a cavity to incubate eggs or feed young and by the sounds of young calling from a nest.

Measurement of Nest Site Characteristics

I measured seven variables describing the characteristics of the stand within a 0.04 ha circular plot centered at each nest. These were 1) habitat (classified as burned or unburned); 2) canopy height (maximum height, measured using a relaskop); 3) basal area of live trees (computed from diameters of all live trees > 8 cm DBH excluding nest tree); 4) shrubcover (estimated percent canopy cover of all woody perennials including trees $\leq$ 8 cm DBH); and 5-7) density of small (< 23 cm DBH), medium (23-38 cm), and large (> 38 cm) snags.

Characteristics recorded for each nest tree included: 1) tree condition (live or dead); 2) diameter (DBH, measured with a diameter tape); 3) height (measured with a relaskop); 4) bark cover (estimated percent of stem covered by bark); 5) top condition (broken or intact); 6) twig condition (most foliage-bearing twigs present or most broken); 7) tree species (Jeffrey pine, lodgepole pine, red fir, white fir, or other).

All binary variables were coded as 0 or 1 for subsequent analyses. Tree species was converted to four dummy binary variables. If the tree was a given species it was assigned a value of 1; otherwise, it was assigned a 0. The fifth group, other, was left out of the analysis to avoid redundancy. In total, then, 17 variables were included in the analysis.

Statistical Analyses

A linear discriminant analysis was performed on all 17 variables (percentages analyzed using an arcsine transformation) using the SPSS (version 8.0) package (Nie et al. 1975). The null hypothesis was that nest sites of all species are equal, that is, the mean discriminant scores do not differ between members of any possible pair of species.

The maximum number of functions derived in a multi-group discriminant analysis is either one less than the number of groups (in this case, bird species) or equal to the number of variables entered into the analysis, whichever is smaller. The first function derived explains the greatest proportion of the total variance, and each additional function explains successively less. There are two approaches in deciding how many of the possible functions to consider. First, one can test the statistical significance of each function by comparing the additional variance explained by that function to an expected value (Klecka 1975, Morrison 1976). Alternatively, one can arbitrarily define a minimum proportion of explained variance and accept only those functions that explain more than that minimum. In this study, I chose the latter approach and considered only those functions explaining 5% or more of the total variance. This alternative allows more powerful planned comparisons of mean discriminant scores among the groups as opposed to post-hoc comparisons which would have been necessary had I used the former approach.

To interpret the biological meaning of each discriminant function, I computed the correlation of each variable with the discriminant score

derived for each function (structure matrix). Variables with the highest correlations were used to interpret functions. Some researchers (e.g., Klecka 1975:443) prefer to use the standardized discriminant function coefficients (pattern matrix) to interpret functions, but these coefficients can be highly unstable.[3]

I used a t-test to compare mean discriminant scores of each pair of species along each of the discriminant axes accepted for analysis. Given S species, there are [S(S-1)]/2 possible pairwise comparisons on each axis. To control the total type I error rate at ≤ 0.05 for all comparisons on each axis, I used Dunn's (1961) procedure for multiple planned comparisons. Assuming equal variance in discriminant scores, the formula used for each pairwise test comparing species i and j on the kth discriminant axis is simply:

$t_k = (\bar{d}_{ik} - \bar{d}_{jk}) / (1/n_i + 1/n_j)^{1/2}$ where $\bar{d}$ is the mean discriminant score and n is the sample size for each species. Values of t were compared to values tabled by Dunn (1961) to accept or reject the null hypothesis of no difference between mean scores at the 0.05 significance level.

The value of t is dependent on sample size as well as the magnitude of the difference between the discriminant scores. Given the same value of $(\bar{d}_{ik} - \bar{d}_{jk})$, a comparison involving two species with large sample sizes may be statistically significant while a comparison of species with smaller sample sizes may not be. Values of t may also be affected by unequal variance in scores among species. To reduce this dependency on sample size and to account for the possibility of unequal variance, I divided each discriminant axis into equal segments, computed the frequency of discriminant scores for each species in each segment, and then computed the similarity of these frequency distributions along each axis using a measure of niche overlap given by Colwell and Futuyma (1971:573, equation 23). Since each discriminant axis is independent, I multiplied the overlap values for each species-pair on each axis to derive an index of total overlap on all axes (May 1975). Cluster analysis (UPGMA, Sneath and Sokal 1973) was used to reveal possible groups of species using nests with similar characteristics based on these total overlap values.

For comparison, I also computed the Euclidian distance (D_{ij}) between each pair of species i and j in discriminant space, using the formula:

$D_{ij} = [\Sigma_k (\bar{d}_{ik} - \bar{d}_{jk})^2]^{1/2}$, where $\bar{d}_{ik}$ is the value of the mean discriminant score on the kth axis for the ith species, and d_{jk} is the value for the jth species on the same axis (D_{ij} is equivalent to the square root of the Mahalanobis distance statistic). Like the overlap values described above, these Euclidian distances are less affected by sample size than the t-test.

To compare the relative importance of the nest tree and nest stand variables as discriminators between nest sites of the species, I performed three additional discriminant analyses. For the first analysis, I included only the seven nest stand variables and computed the proportion of the total variance explained by all functions (sum of squares between groups divided by the total sum of squares). Second, I used a stepwise analysis to determine the best seven of the ten nest tree variables and calculated the proportion of the total variance explained by these seven variables. Third, I used a stepwise analysis including both the nest tree and nest stand variables and calculated the proportion of the total variance explained by the first seven variables entered of either type. I then compared these values of explained variance from each analysis to assess the relative importance of the tree and stand variables.

RESULTS AND DISCUSSION

Interpretation of Functions

I located a total of 306 nests of 9 excavator species (table 1). These species formed nine groups resulting in eight discriminant functions which explained 83% of the total variance in nest site characteristics (fig. 1). Each of functions 4 through 8 explained less than 5% of the variance not accounted for by previous functions and was not considered in subsequent analyses. The first three functions explained 55%, 11%, and 7%, respectively, totaling 75% of the variance.

The first function was correlated most highly with canopy height, live tree basal area, and burned vs. unburned habitat (table 2). This function clearly was associated with nest stand variables and separates species nesting in unburned forest stands (red-breasted nuthatch and sapsuckers) from those nesting in burned stands (pygmy nuthatch, and Lewis and white-headed woodpeckers) (fig. 2A). The three remaining species nest in both burned and unburned habitats and their mean discriminant scores were located at intermediate locations on the first discriminant axis (fig. 2A).

The second discriminant function was most correlated with nest tree variables, particularly species (the two fir categories) and tree size (height and diameter) (table 2). Species nesting in red fir and smaller size trees had the lowest discriminant scores (black-backed and white-headed woodpeckers) and species nesting in larger size, white fir trees had the highest scores (Lewis

[3]Personal communication with L.A. Marascuilo, statistics Professor, Department of Education, University of California, Berkeley.

Table 1. Bird species (hole excavators) nesting in the Sagehen Creek study area.

Bird species	Code	Sample size
Common flicker (*Colaptes auratus*)	CF	68
Lewis woodpecker (*Asyndesmus lewis*)	LW	37
Red-breasted sapsucker (*Sphyrapicus ruber daggetti*)	RS	50
Williamson sapsucker (*Sphyrapicus thyroideus*)	WS	50
Hairy woodpecker (*Dendrocopos villosus*)	HW	23
White-headed woodpecker (*Dendrocopos albolarvatus*)	WW	12
Black-backed-three-toed woodpecker (*Picoides arcticus*)	BW	8
Red-breasted nuthatch (*Sitta canadensis*)	RN	30
Pygmy nuthatch (*Sitta pygmaea*)	PN	28
TOTAL		306

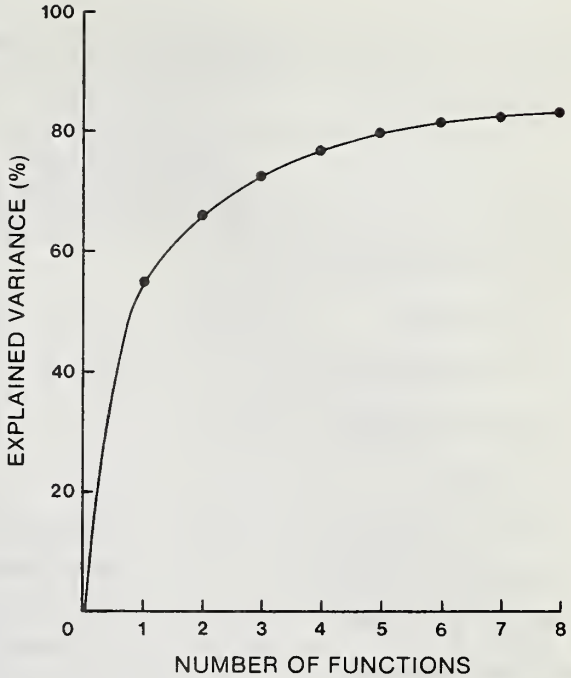


Figure 1. Cumulative explained variance (% of total variance) in relation to the number of discriminant functions considered.

woodpecker and red-breasted-sapsucker) (fig. 2B). The mean scores of the other five species differed very little on this axis.

The third discriminant function was also associated with variables describing the nest tree (table 2) and it separated the white-headed, hairy, and black-backed woodpeckers from the remaining species (fig. 2C). The white-headed woodpecker nests in large diameter, broken-topped trees while the black-backed and hairy woodpeckers nest in smaller diameter, intact-topped trees. As with the second function, the remaining species have very similar mean scores on this axis.

Pairwise Comparisons

Results of the 36 t-tests comparing mean discriminant scores among all possible pairs of species on each discriminant axis are summarized in figure 3. The null hypothesis of no difference between mean scores was rejected on at least one axis for all but three comparisons (RN vs. WS, BW vs. HW, WS vs. RS). These results might be used to suggest that the black-backed and hairy woodpeckers could be grouped for management purposes because their nest site characteristics were statistically indistinguishable. Similarly, the two sapsuckers could also be grouped. The pygmy nuthatch is more problematic since its nest sites are similar to those of the Williamson sapsucker but not to those of the red-breasted sapsucker; it might be better, therefore, not to group pygmy nuthatches with either species.

Because all discriminant axes are orthogonal (independent), a significant difference between two species' centroids on any one axis is sufficient to conclude that the two species use statistically distinct nest sites. Given that the null hypothesis is true and that the probability of at least one significant difference on each axis is 0.05, the probability of no difference on any axis is $(1.00 - 0.05)^3 = 0.86$. Thus, for 36 comparisons one could expect approximately $36(0.86) = 31$ to be nonsignificant. Clearly, since only three comparisons were actually nonsignificant, the null hypothesis is not true. It is unclear, however, whether the lack of a significant difference among these three pairs reflects their inherent ecological similarity or a chance event.

Interpretation of these pairwise comparisons is further complicated by the fact that the comparisons along any one axis are not independent. When significant differences do occur among comparisons of this type, they tend to occur in "bunches" (Lindman 1974:82). Therefore,

Table 2. Pooled within-groups correlations between discriminant functions and discriminating variables.

Variable	Correlation with Discriminant Function: 1	2	3
Nest Stand Variables			
Canopy height	0.83	0.12	-0.03
Live tree basal area	0.70	0.07	-0.16
Burned or unburned	-0.59	0.10	-0.03
Shrub cover	-0.26	-0.19	0.12
Snags < 23 cm DBH	0.02	-0.18	0.05
Snags 23-38 cm DBH	-0.14	-0.06	0.05
Snags > 38 cm DBH	-0.15	0.01	-0.07
Nest Tree Variables			
Height	0.51	0.43	0.20
Diameter	0.31	0.32	-0.53
Foliage-bearing twigs	0.52	0.19	0.14
Bark cover	0.29	0.17	-0.27
Top condition	0.23	-0.17	0.38
Jeffrey pine	-0.26	0.04	-0.21
Lodgepole pine	0.21	-0.06	0.31
White fir	-0.08	0.28	0.00
Red fir	0.08	-0.47	-0.15
Tree condition	-0.44	-0.10	0.01

one cannot predict, a priori, the expected number of significant differences on one, two, or three discriminant axes. For these reasons, pairwise comparisons of species' mean discriminant scores appear to have limited utility in assessing habitat similarity among the species. Two other measures, involving overlap of score distributions and distance between mean scores, seem better suited for such a purpose.

Overlap of Discriminant Scores

The range and standard deviation of discriminant scores on the discriminant axes show considerable overlap among species (fig. 2). Two species may have statistically different mean scores but their score distributions could overlap to such an extent that they might justifiably be grouped for management purposes. To examine such a possibility I computed overlaps between all pairs of species on each of the three discriminant axes and multiplied the three overlap values for each species-pair to derive an index of nest site similarity (table 3). The greatest overlap values are found between the two sapsuckers (0.42) and between the common flicker and pygmy nuthatch (0.40).

I used a cluster analysis based on this similarity matrix (overlap values) to produce a dendrogram (fig. 4) revealing patterns of nest site similarity among the bird species. At the 0.4 level, two groups are recognized, one composed of the two sapsuckers and another containing the pygmy nuthatch and common flicker. At the 0.2 level, the Lewis woodpecker links with the flicker-pygmy group and the red-breasted nuthatch joins the sapsucker group. At this level of similarity, these groups can be indentified as burn and unburned specialists, respectively. The remaining species link at successively lower overlap values. These results suggest that nest sites of the two sapsuckers may be similar enough to permit their management in common and, likewise, those of the pygmy nuthatch and common flicker. The pairwise t-test also identified the sapsucker group, but the t-tests indicated the hairy and black-backed woodpeckers as a second possible group rather than the nuthatch and flicker.

Euclidian Distance

Another measure of overall nest site similarity is the Euclidian distance between species' mean discriminant scores. These values (table 3) measure the separation of the species in discriminant space, and may be less affected by sample size than the overlap values. For example, the overlap of the black-backed woodpecker with the other species ranges from 0.01 to 0.02 (table 3). These values are lower than the average of any other species, probably reflecting the small sample size of black-backed nests (n = 8). In contrast, the Euclidian distances between black-backed nests and those of the other species range from 1.30 to 3.24 and these values are spread throughout the range exhibited by the other species-pairs.

Euclidian distances, however, do not directly convey any information about the dispersion of each species about its centroid. The overlap values do convey such information, since it is precisely the relative patterns of dispersion of any two species that define their overlap. But, a linear regression of discriminant score overlap with Euclidian distance between mean scores in three-dimensional discriminant space showed that the two measures were highly correlated ($r = -0.89$, $P < 0.001$, excluding the black-backed woodpecker). Thus, overlap can be predicted from Euclidian distance. Given this relationship, and

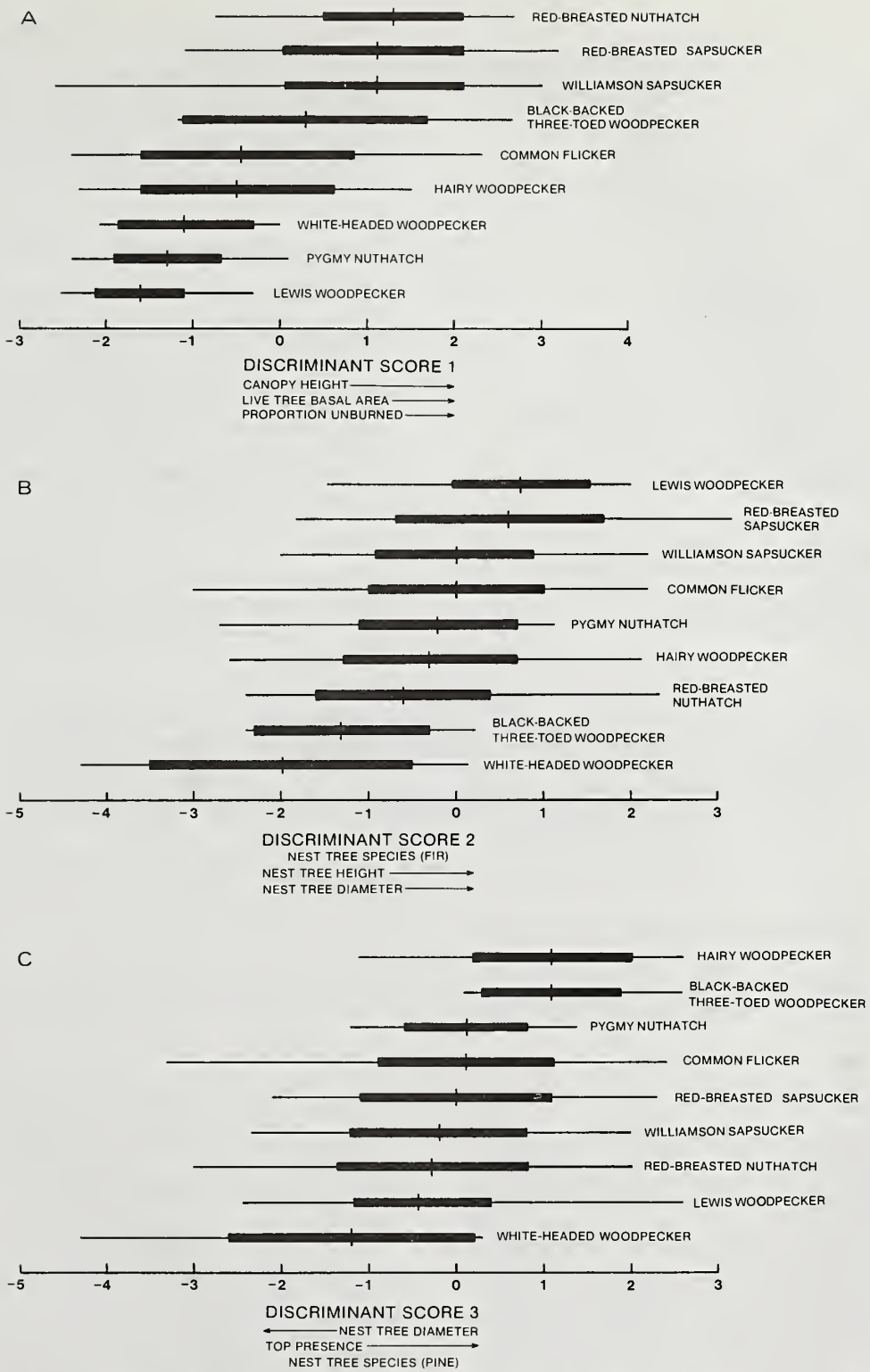


Figure 2. Mean (vertical lines), standard deviation (heavy bars), and range (horizontal lines) of discriminant scores of each bird species on the first (A), second (B), and third (C) discriminant functions. Arrows indicate direction of increased values of variables most highly correlated with the discriminant scores on each function.

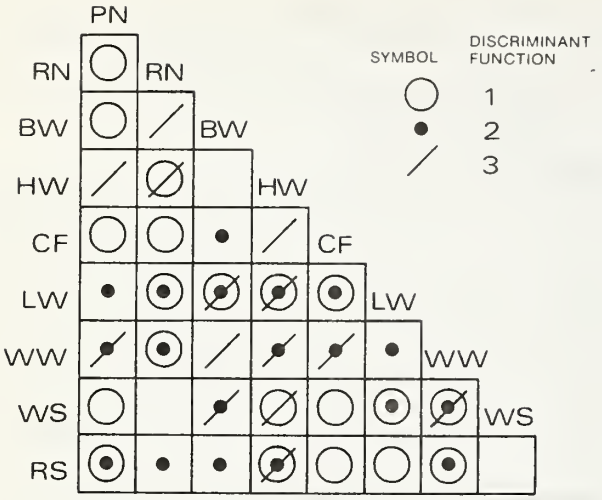


Figure 3. Results of t-tests comparing mean discriminant scores among all possible pairs of species on each function. Symbols denote significant differences (see text). Species codes are given in table 1.

the observation that Euclidian distances are apparently less dependent on sample size than the t-tests or overlap values, I recommend using Euclidian distance as a measure of similarity when sample sizes of some groups are small. When sample sizes are large, the overlap value is preferred since it is directly interpretable as a measure of shared discriminant score distribution.

Relative Importance of Stand and Tree Variables

Figure 3 shows that eight pairs of species differ only on the basis of stand characteristics (significant differences occur only on the first discriminant axis). These species nest in similar trees located within otherwise different stands. In contrast, 13 species-pairs nest in trees with significantly different characteristics (significant differences occur only on functions 2 and/or 3). They nest in differing trees located in otherwise similar stands. The remaining 12 pairs of species nest in both different trees and stands.

Three additional discriminant analyses, each using seven variables, were used to assess the relative discriminating power of the tree and

Table 3. Relative similarity of excavator nest sites. Values to right of diagonals are the overlaps of species' discriminant score distributions (larger values indicate more similar nest sites). Values to left of diagonals are Euclidean distances between species' mean discriminant scores in three-dimensional discriminant space (larger values indicate less similar nest sites).

Bird species[1]	PN	RN	BW	HW	CF	LW	WW	WS	RS
PN	----	0.02	0.03	0.19	0.40	0.22	0.11	0.05	0.09
RN	2.72	----	0.02	0.06	0.17	0.01	0.01	0.24	0.21
BW	2.21	1.90	----	0.02	0.02	0.01	0.02	0.01	0.02
HW	1.30	2.36	1.33	----	0.22	0.04	0.02	0.10	0.13
CF	0.91	1.93	1.85	1.08	----	0.26	0.09	0.28	0.32
LW	1.15	3.24	3.24	2.20	1.48	----	0.01	0.02	0.04
WW	2.23	2.91	2.75	2.93	2.48	2.87	----	0.02	0.02
WS	2.47	0.69	2.09	2.15	1.59	2.82	3.12	----	0.42
RS	2.59	1.25	2.40	2.21	1.68	2.79	3.59	0.60	----

[1]See table 1 for bird species codes.

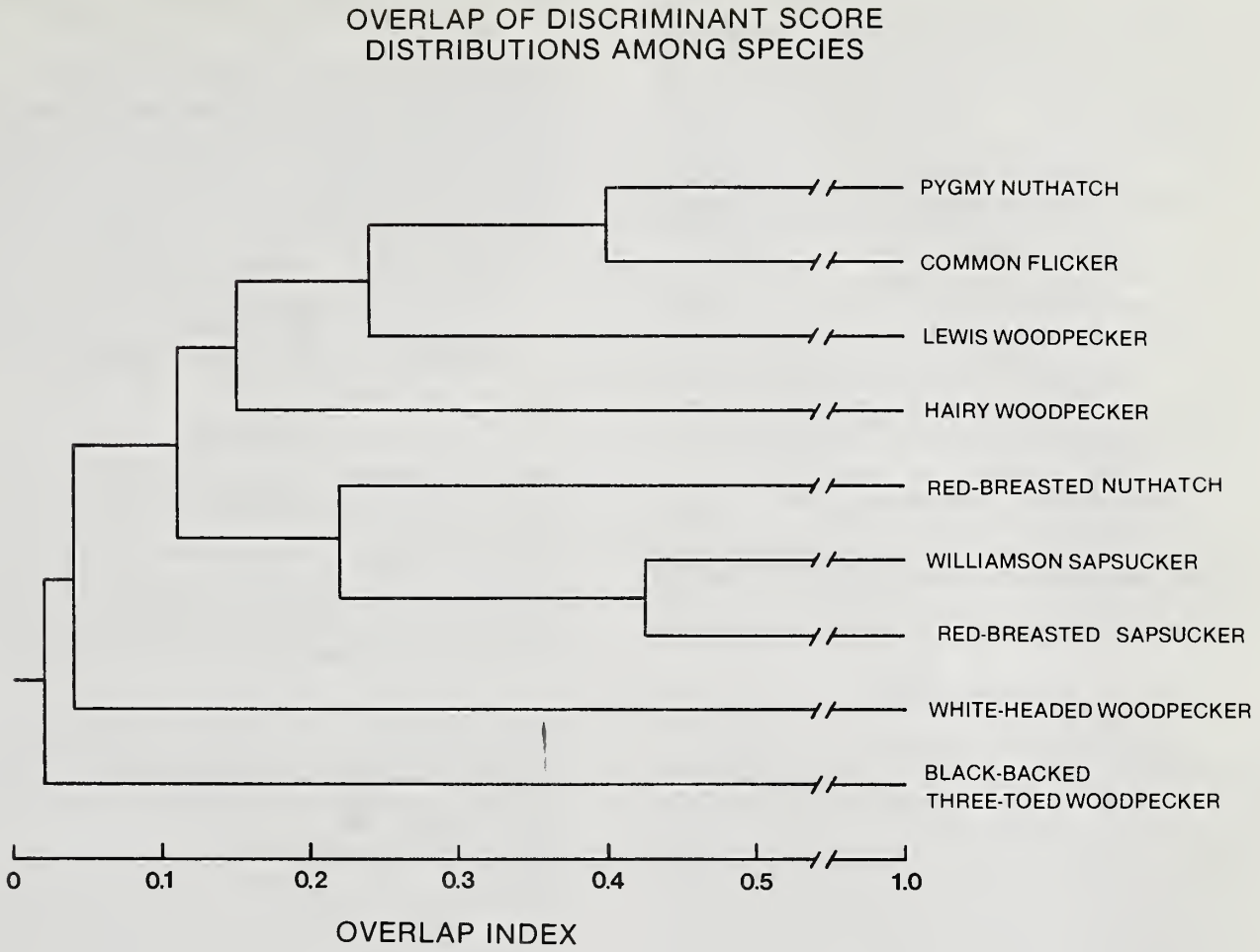


Figure 4. Dendrogram showing similarity of nest sites of bird species based on cluster analysis using overlap of discriminant score distributions along three discriminant axes. Larger index values indicate more similar nest sites.

stand variables (see methods). When only the seven stand variables were included in the analysis, the variance explained by the set of discriminant functions was 60% of the total variance in the system. The variance explained using only the best seven nest tree variables (lodgepole pine, white fir, and foliage bearing twigs were excluded) is slightly higher, 63%. Finally, the optimum set of both nest and stand variables (canopy height, tree height, tree diameter, red fir, live tree basal area, top condition, and tree condition listed in order of entry in stepwise analysis) explained 74%. While the nest tree variables are slightly better than the stand variables as discriminators between species' nest sites, it is apparent that both types of variables should be included to adequately characterize nest sites. Consideration of both types resulted in a substantially greater separation of species' nest sites.

CONCLUSION

Discriminant analysis is a useful method for quantifying the similarity of habitat characteristics among wildlife species. It offers several options including planned or post-hoc comparisons of mean scores, analysis of score distributions in discriminant space, and measures of the separation of species such as Euclidian distance. Of these options, Euclidian distance appears to have the greatest generality since it is affected less by variation in sample size among the species analyzed.

Using discriminant analysis to compare the characteristics of the nest sites of nine sympatric primary cavity nesting bird species, I demonstrated that both nest tree and nest stand characteristics should be included in the suite of variables used by managers to prescribe management

practices to provide the habitat requirements of these species. Five species used distinct nest sites and require unique management prescriptions, but the four other species can be combined for management purposes into two groups, each containing pairs of species using very similar nest sites.

As a final note, researchers (and managers) must recognize that interspecific comparisons of habitat characteristics are not sufficient to define the most important variables involved in habitat selection. For example, nests of all species considered in this study were surrounded by a relatively large number of snags greater than 23 cm DBH. But because the magnitude of this variable does not vary significantly among species, snag density is not a good discriminator between species' nest sites. When snag density is compared between a species' nest sites and a set of random plots, however, it becomes one of the best discriminators (Raphael 1980). Thus, variables important in habitat selection should be identified through a series of single species analyses comparing use and availability of habitat components rather than by comparing habitat characteristics among all species of interest.

ACKNOWLEDGMENTS

I thank the U.S. Forest Service, Region 5, and my wife, Susan, for financial support. Marshall White and Vernon Hawthorne provided field facilities at the Sagehen Creek Field Station. Computer time was provided by grants from the Department of Forestry and Resource Management, U.C. Berkeley. I thank David Capen and Jake Rice for their useful technical and editorial review of the manuscript. I also express my appreciation to Nobu Asami for typing several drafts of the manuscript.

LITERATURE CITED

Colwell, R.K., and D.J. Futuyma. 1971. On the measurement of niche breadth and overlap. Ecology 52:567-576.

Dunn, O.J. 1961. Multiple comparisons among means. Journal of the American Statistical Association 56:52-64.

Klecka, W.R. 1975. Discriminant analysis. p. 434-467. In Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Bent. Statistical package for the social sciences. Second edition. 675 p. McGraw-Hill, New York, N.Y.

Lindman, H.R. 1974. Analysis of variance in complex experimental designs. 352 p. W.H. Freeman, San Francisco, Calif.

May, R.M. 1975. Some notes on estimating the competition matrix, α. Ecology 56:737-741.

Morrison, D.F. 1976. Multivariate statistical methods. Second edition. 415 p. McGraw-Hill, New York, N.Y.

Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Bent. 1975. Statistical package for the social sciences. Second edition. 675 p. McGraw-Hill, New York, N.Y.

Raphael, M.G. 1980. Utilization of standing dead trees by breeding birds at Sagehen Creek, California. Ph.D. dissertation. 195 p. University of California, Berkeley.

Raphael, M.G., and M. White. 1978. Snags, wildlife, and forest management in the Sierra Nevada. Cal-Neva Wildlife 1978:23-41.

Sneath, P.H.A. and R.R. Sokal. 1973. Numerical taxonomy. 573 p. W.H. Freeman, San Francisco, Calif.

Thomas, J.W., R.G. Anderson, C. Maser, and E.L. Bull. 1979. Snags. p. 60-67. In Thomas, J.W., editor. Wildlife habitats in managed forests--the Blue Mountains of Oregon and Washington. USDA Forest Service, Agriculture Handbook No. 533. 512 p. Washington, D.C.

DISCUSSION

DONALD McCRIMMON: Please reiterate the method used to interpret discriminant axes.

MARTIN RAPHAEL: I calculated the correlations between each original variable and the discriminant score for each function (structure matrix) and interpreted the functions by considering the most highly correlated variables.

PAUL GEISSLER: I think it would be more appropriate to use a multivariate analysis of variance (MANOVA) to test for differences among species, possibly with orthorgonal contrasts based on taxonomic relationships. The criteria in MANOVA are functions of the eigenvalues associated with the eigenvectors which define the discriminant functions. It has been suggested that plots of the treatment means in the space of the discrimination functions be used to interpret MANOVA results.

MARTIN RAPHAEL: I believe that a one-way MANOVA and the multi-group discriminant analysis are the same procedure. I would not want to limit the contrasts to taxonomic relations. For example, I found that the common flicker and pygmy nuthatch--two very distantly related species--were highly similar in nest site selection.

KEN WILLIAMS: Did you check for covariance hetergeneity? Without this characteristic it is difficult to test your stated hypothesis. Also, it assures some degree of overlap distortion as displayed by the discriminant functions.

MARTIN RAPHAEL: I did test for variance/covariance heterogeneity, and the variance/covariance matrices were not equal. I could find no method to assess the consequences of the failure to meet this assumption, and decided to proceed with the analysis since some authors state that the procedure is robust against such a failure. My understanding is that this failure has more severe implications on the interpretation of discriminant coefficients. It is quite possible that distortions occurred in the pattern of overlaps among the species; I would ask how severe such distortion might be: are they so

severe that the results are wrong? No one, as yet, can answer the latter question.

LESLIE MARCUS: I suggest that you plot points or bivariate ellipse projections into canonical variate space, to see how covariance-structure looked in bivariate space. You could use the term "Mahalanobis distance" rather than "Euclidean distance in discriminant space"--as one always knows what it means.

MARTIN RAPHAEL: To answer the first point, I cannot visualize the advantage of plotting such projections since any distortions caused by the covariance structure would still affect the bivariate axes. I would like to know more about such a technique if it really would allow an evaluation of distortions. On the second point, I suppose it is a matter of personal taste since the two measures are so closely related. It would be useful to agree on the use of either measure to facilitate communications.

ROBUST CANONICAL CORRELATION OF SAGE GROUSE HABITAT[1]

Mark S. Boyce[2]

Abstract.--The distribution patterns of sage grouse (Centrocercus urophasianus Bonaparte) were studied on Atlantic Richfield Company's Coal Creek coal surface mine site in northeastern Wyoming. Fecal droppings were periodically removed from randomly located belt transects within a 36 km² study area. Using multiple regression and canonical correlation the seasonal distribution of droppings was related to several habitat variables which were hypothesized to be important to sage grouse. By canonical correlation, approximately 90% of the variance in a fecal count variate can be attributed to a habitat variate. The correlations between each variate and the original variables were found to be very stable when habitat variables were rotated. Similarly, the robustness of the loadings was confirmed by systematically eliminating cases (transects) from the analysis. Although canonical correlation provides little insight here over that obtained from multiple regression analysis, it does provide a succinct summary of a large and complex set of interrelationships among variables.

Key words: Canonical correlation; habitat; mine reclamation; multiple regression; patchiness; sage grouse; Wyoming.

INTRODUCTION

Applications of canonical correlation in ecology are relatively uncommon (Smith 1980). This may be due in part to unstable results or difficult interpretation (Gauch and Wentworth 1976, Johnston[3]). These problems are often attributable to assumption violations in the data, e.g., variables which are not normally distributed (Harris 1975), multicollinearity among variables (Cohen et al. 1979), or nonlinearity (Gauch and Wentworth 1976). However, if these assumptions can be approximated, the potential is great for application of canonical correlation in habitat studies. Canonical correlation searches for the maximum correlation between linear combinations of two sets of variables; for example, one set of variables may consist of measures of organism distribution and/or abundance, and another set may consist of variables characterizing habitat.

In this paper I describe an application of canonical correlation to an analysis of sage grouse habitat. One set of variables consists of the number of sage grouse fecal droppings deposited in different seasons, whereas the second set of variables consists of selected habitat variables which were hypothesized to be important factors influencing sage grouse distribution. Correlations between the original variables and the derived distribution and habitat variates were found to be robust (contra Cohen et al. 1979), even though the sample size was relatively small

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Assistant Professor, Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071.

[3]Personal communication with R.F. Johnston, Professor, University of Kansas.

(n = 20). I attribute the robustness of this application of canonical correlation to thorough screening for nonlinearity and normality followed by transformation or elimination of variables before conducting the analysis.

METHODS AND MATERIALS

The study area was a 36 km² coal lease (T46N, R70W) in Campbell County, Wyoming which is scheduled for development as a surface mine. An objective of this research was to characterize the habitat components which are most important to sage grouse and to develop guidelines for reclamation of mined lands for sage grouse.

Sage grouse fecal droppings are easily identified and may last for up to 3 years before deteriorating. A sampling scheme was designed to assess seasonal distribution of sage grouse by periodically checking permanent belt transects for fresh droppings, and removing all droppings deposited since transects were last searched. Twenty 2 m x 1000 m belt transects were randomly located on the study area, and checked at irregular intervals.

An independent consulting team was employed to map and characterize all vegetation types on the study area (Keammerer and Keammerer 1975). In addition, detailed data were collected on the composition of sagebrush density, height and cover along each transect by employing systematic point-quarter procedures (Seber 1973).

The total number of fecal pellets found during a particular season was summed over the three years of our study, yielding three seasonal distribution variables, i.e., number of droppings found in 1) winter, 2) spring, and 3) summer through early fall. Habitat use patterns are known to be very similar in summer and early fall in non-migratory sage grouse populations (Patterson 1952). Habitat and distribution variables employed in this analysis are listed in table 1.

RESULTS

An average of 360 droppings was found on each transect. All variables were screened for skewness and kurtosis, and transformations conducted where appropriate. Seven variables were either eliminated or combined with other variables because their distribution contained an excessive proportion of zeros. Next, because of the sensitivity of canonical correlation to non-linearity (Gauch and Wentworth 1976), bivariate plots were constructed between the distribution variables and each habitat variable. When nonlinear patterns appeared, the respective habitat variable was eliminated from the analysis. For example, the total droppings vs. the area within each belt transect in the big sagebrush habitat type appears nonlinear (fig. 1). The availability of at least some big sagebrush is

Table 1. Reduced set of variables and appropriate transformations for variables used in this analysis.

Distribution variables

TOTAL	= $\log_{10}$ (total droppings found on each transect over 3 years, plus 1.0).
SUMMER-FALL	= $\log_{10}$ (sum over 3 years of all droppings which accumulated between June and October, plus 1.0).
WINTER	= $\log_{10}$ (droppings accumulated during winter months, plus 1.0).
SPRING	= $\log_{10}$ (sum over 2 years of droppings accumulated during spring months through mid-May, plus 1.0).
DIST	= fecal droppings canonical variate.

Habitat variables

LEK	= distance between the nearest sage grouse strutting ground (lek) and the closest point on each belt transect in km.
FORBS	= summed area within each belt transect within forb-producing habitats, including riparian areas, in m².
DIV	= habitat diversity = $-\Sigma p_i \log p_i$ where p_i is the proportion of the ith habitat type within each belt transect.
PATCH	= patchiness index equal to the number of times the belt transect crosses from one habitat type to another, plus 1.0.
SPRAY	= $\log_{10}$ (area within belt transect of big sagebrush habitat which had been sprayed with 2,4-D, plus 1.0).
COVER	= average sagebrush cover along transect estimated with 21 point-quarter measurements (Seber 1973).
VACOV	= standard deviation of COVER.
DENS	= average density of sagebrush plants per m² estimated from 21 point-quarter measurements (Seber 1973).
VADEN	= standard deviation of DENS.
SAGE	= area within belt transect of big sagebrush habitat type in m².
HABITAT	= canonical variate comprised of habitat variables.

essential, and transects with no big sagebrush have no sage grouse utilization. However, large homogeneous stands of sagebrush tend to have low forb availability and are not preferred habitat. Therefore, SAGE was eliminated from further analysis.

A multiple regression analysis was conducted using each distribution variable as a dependent variable. This permitted explicit hypothesis testing regarding the importance of various habitat variables. Several interesting patterns emerge; for example, habitat patchiness and the proportion of forb-producing habitats were both positively correlated with the number of droppings found on each belt transect (fig. 2). Different habitat variables are important at different times of the year due to the seasonal changes in food habits and behavior of the grouse. A few of the more interesting multiple regression models are summarized in table 2.

In an attempt to summarize the complex interrelationships between the seasonal distribution patterns and habitat, I conducted a canonical correlation analysis of the "distribution" set of variables and a selected set of habitat variables. Due to the small sample size (n = 20), I began with a small number of variables: three distribution variables, and three of the most important habitat variables. As presented in figure 3A, a habitat variate accounts for almost 90% of the variance in a distribution variate and the relationship is highly significant ($P << 0.001$). Furthermore the correlations between the first pair of canonical variates and the original variables are all intuitively satisfying and parallel the results of multiple

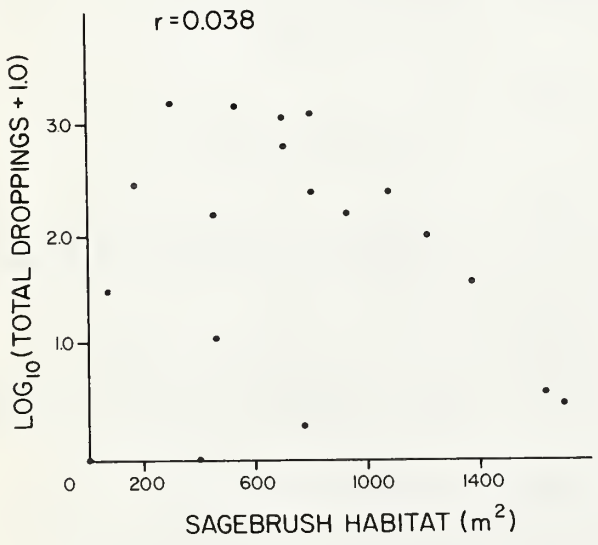


Figure 1. Total droppings over three years plotted as a function of the area within each belt transect in the big sagebrush habitat type.

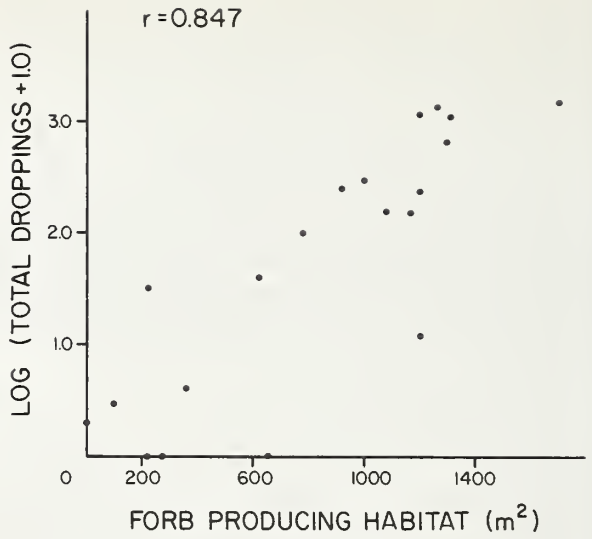


Figure 2. Total droppings over there years as a function of the area within each belt transect in forb-producing habitat types.

regression. The second canonical correlation was not statistically significant ($P < 0.05$), and loadings of the second pair of variates were not readily interpretable.

Since the 3 x 3 analysis worked so well, the number of habitat variables in the model was increased to assess whether a better synopsis of the interrelationships between habitat and distribution could be achieved. The analysis was attempted with five habitat variables (fig. 3B) and then with nine habitat variables (fig. 3C). The high stability of the loadings was impressive, i.e., the correlations between the first pair of canonical variates and the original variables remained relatively constant as additional habitat variables were added to the analysis. Even more exciting, however, the loadings were intuitively reasonable, and provided an exceptionally good representation of the interrelationships between variables which I had discovered through a tedious and detailed multiple regression analysis. All distribution variables loaded positively into a distribution variate, thus clearly the distribution variate reflected intensity of use by sage grouse. Correlations between original variables and the habitat variate all reflected quality of sage grouse habitat, i.e., variables which were positively correlated tend to be preferred habitat components and variables which were inversely correlated with the habitat variate were usually inversely correlated with the number of droppings found on transects at various times of year.

To test the robustness of the canonical correlations and variable loadings, the 3 x 9 analysis was rerun 20 times, each time leaving out one case (transect) from the analysis (similar to

Table 2. A sample of multiple regression models with droppings counts as dependent variables and habitat attributes as independent variables. All models listed account for a significant proportion of the variance in dependent variables ($P < 0.05$).

SPRING = 2.674 - 0.452(LEK) - 0.056(SPRAY)	R = 0.780
SPRING = 1.603 - 0.359(LEK) + 0.002(FORBS)	R = 0.841
SPRING = 0.002(FORBS) + 10.6(COVER) - 0.413	R = 0.818
SPRING = 3.367(DIV) + 13.0(COVER) - 0.686	R = 0.745
SPRING = 0.936 - 0.251(LEK) + 0.002(FORBS) + 5.8(COVER)	R = 0.864
SPRING = 2.706 - 0.481(LEK)	R = 0.779
SPRING = 0.003(FORBS) - 0.163	R = 0.680
SPRING = 0.313 + 14.56(COVER)	R = 0.671
SUMMER-FALL = 0.826 + 0.002(FORBS) - 0.414(SPRAY)	R = 0.896
SUMMER-FALL = 0.467 + 0.116(PATCH) + 0.002(FORBS) - 0.416(SPRAY)	R = 0.912
SUMMER-FALL = 0.664 + 2.3(DIV) - 0.265(LEK) + 0.002(FORBS)	R = 0.879
SUMMER-FALL = 0.812 + 2.28(DIV) - 0.133(LEK) + 0.001(FORBS) - 0.325(SPRAY)	R = 0.923
SUMMER-FALL = 0.994 + 3.54(DIV) - 0.165(LEK) - 0.371(SPRAY)	R = 0.910
SUMMER-FALL = 0.623 + 3.56(DIV) - 0.528(SPRAY)	R = 0.887
SUMMER-FALL = 0.995 - 0.241(LEK) + 0.002(FORBS)	R = 0.861
SUMMER-FALL = 1.88 - 0.624(SPRAY)	R = 0.810
SUMMER-FALL = 0.003(FORBS) - 0.191	R = 0.782
WINTER = 1.747 - 2.5(DIV) + 0.002(FORBS) - 0.423(SPRAY) - 0.238(VADEN)	R = 0.680
WINTER = 1.234 + 0.001(FORBS) - 0.42(SPRAY) - 0.216(VADEN)	R = 0.649
WINTER = 1.935 - 0.555(SPRAY) - 0.207(VADEN)	R = 0.583
WINTER = 1.184 - 0.314(SPRAY)	R = 0.437
TOTAL = 1.42 + 4.798(DIV) - 0.374(LEK)	R = 0.823
TOTAL = 0.718 + 0.003(FORBS) - 0.26(SPRAY)	R = 0.881
TOTAL = 2.415 - 0.598(SPRAY)	R = 0.691

a jackknife procedure). Results for the first pair of canonical variates were impressively stable, with one exception. When transect number 16 was eliminated from the analysis, the loadings for the first pair of canonical variates were not interpretable (table 3). However, upon inspection I discovered that the second pair of canonical variates were loaded precisely in the same pattern as the first pair in all of the other rotations. Thus, the same patterns were present in all trials, but for some reason the second orthogonal pair of variates switched with the first when transect number 16 was deleted from the analysis. The second canonical correlation was statistically significant ($P < 0.05$) in this case whereas it was not in any of the other analyses.

The results of the robustness assessment are summarized in table 4 where the means and standard deviations of the canonical correlations and variable loadings are presented for 20 rotations. For the analysis where transect 16 was eliminated, the second pair of canonical variates was used, but for all other cases the results were averaged over the first pair of canonical variates.

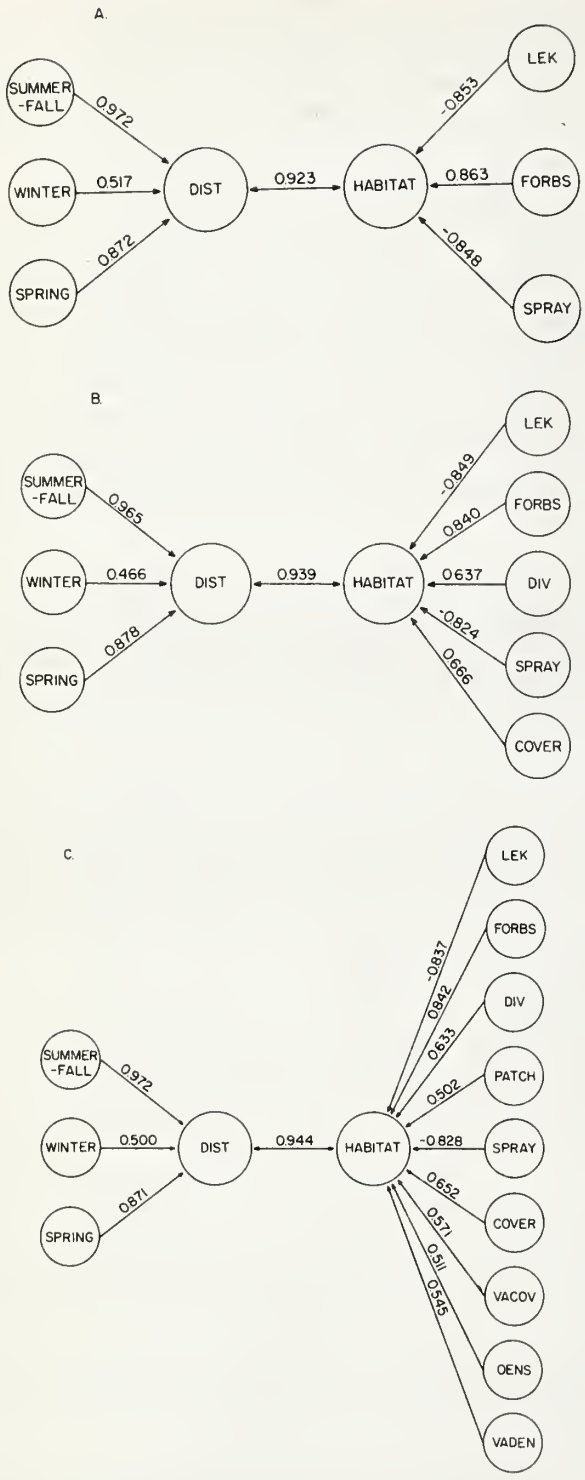


Figure 3. Path diagram of the first pair of canonical variates with three distribution variables and three (A), five (B), and nine (C) habitat variables. The significance levels based upon Bartlett's test are P = 0.00003, 0.00006 and 0.004 for the 3 x 3, 3 x 5, and 3 x 9 analyses respectively. Definitions of variables are listed in table 1.

To further assess the robustness of the 3 x 9 canonical correlation analysis, I systematically eliminated one of nine habitat variables from the analysis. The canonical correlations and variable loadings proved very stable for the first pair of canonical variates. In this experiment as well as during the rotation of cases, loadings into the second pair of canonical variates fluctuated dramatically. As can be seen in table 4, however, variation in canonical correlations and variable loadings for the first pair of canonical variates is impressively small.

The importance of distance to a lek is not clear from these results. It seems plausible that leks are located in good habitats but this may obscure results since grouse social behavior may override habitat quality in determining their distribution. My first attempt to better understand the importance of leks entailed using regression analysis to remove variation attributable to LEK from each of three distribution variables. Then canonical correlation analysis was conducted between the residuals of the distribution variables and the eight original habitat variables with LEK (sensu Boyce 1978). However, the canonical correlation was not statistically significant, and variable loadings were not interpretable. Further multiple regression analysis of the three distribution variables was conducted by studying residuals after removing variation attributable to the area in sprayed sagebrush habitat (SPRAY). The partial correlation between the fecal dropping variables and LEK was discovered to be significant only in the springtime ($r = -0.639$, $P < 0.01$). Since the leks are only actively used during March, April, and May, it seems plausible that only in spring should there be any unique contribution to distribution determined by the proximity to a lek.

DISCUSSION

Sample size guidelines are lacking for applications of many multivariate procedures. Although Bartlett's test for canonical correlation allows one to assess the overall significance of the relationship between two sets of variables, there is presently no means by which significance of individual variable loadings can be evaluated. Nevertheless, the most useful applications of procedures such as canonical correlation may be heuristic ones, thus inference procedures may not be necessary or even desirable. However, assessing the goodness of such descriptive tools does seem important, and robust procedures such as jackknifing provide such an assessment. If canonical correlations and variable loadings prove to be robust, as they were in this study, the extremely large sample sizes recommended by Thorndike (1978) should not be necessary. The impracticality of collecting extremely large data sets in wildife habitat studies should not persuade the investigator that applications of multivariate procedures, such as canonical correlation, are necessarily inappropriate.

Table 3. Correlation between original variables and the first two pairs of derived canonical variates when transect number 16 was eliminated from the analysis. Note the similarity between figure 3C and the loadings for the second pair of canonical variates. The probability levels are from Bartlett's test.

	Distribution variate		Habitat variate	
First Canonical	SUMMER-FALL	-0.542	LEK	0.090
r = 0.973	WINTER	-0.548	FORBS	-0.566
(P = 0.0002)	SPRING	-0.040	DIV	-0.229
			PATCH	-0.052
			SPRAY	0.544
			COVER	0.053
			VACOV	-0.041
			DENS	0.072
			VADEN	-0.082
Second Canonical	SUMMER-FALL	0.827	LEK	-0.858
r = 0.931	WINTER	0.424	FORBS	0.647
(P = 0.048)	SPRING	0.924	DIV	0.589
			PATCH	0.573
			SPRAY	-0.655
			COVER	0.707
			VACOV	0.567
			DENS	0.551
			VADEN	0.425

Cohen et al. (1979) claim that canonical weights (loadings) are necessarily unstable in canonical correlation because each variable accounts for a unique portion of the variability in the other set of variables. However, we have seen in this analysis that unstable loadings or "bouncing betas" need not occur and that removal or addition of variables need not influence the qualitative interpretation. It seems more likely that nonlinearity or multicollinearity are behind many applications where "bouncing betas" appear to be a serious problem (see Gauch and Wentworth 1976).

The results depicted in figure 3 confirm many of the subjective impressions which were developed based upon field obsevations of the major habitat components important to sage grouse. During winter, sage grouse feed exclusively on leaves of big sagebrush (Artemisia tridentata). However, between April and October, forbs such as dandelion (Taraxacum officianale), curly-cup gumweed (Grindelia squarrosa), sweet clover (Melilotus officianalis), false dandelion (Agoseris glauca), alfalfa (Medicago sativa), and salsify (Tragopogon dubius) constitute a major portion of the diet for both young and adult birds (Patterson 1952). Although sagebrush provides important visual cover from aerial predators, sagebrush plants compete with forbs for nutrients and water. Thus optimal habitats are often patchy where forbs and sagebrush cover occur in close proximity.

Each of the variables which we selected to measure habitat patchiness (DIV, PATCH, VACOV, VADEN) is positively correlated with the number of droppings found on the transects. Similarly, these same variables are positively correlated with the "habitat quality" variate. SPRAY is clearly an important variable which is negatively loaded into the "habitat quality" variate. Since spraying with herbicides such as 2,4-D kills both forbs and sagebrush, it is easy to appreciate the devastating impact of herbicide spraying on the use of areas by sage grouse.

It is important to caution against cavelier interpretations of canonical correlation analysis. As noted earlier, explicit hypothesis testing regarding any single variable is not possible with canonical correlation but may be accomplished with linear regression techniques. Some patterns can be masked by simple reliance upon the overall trends summarized by canonical correlation as I demonstrated above for LEK. Also, to achieve a robust canonical model, variables must often be eliminated for statistical reasons even though some of these variables may be biologically important, e.g., SAGE. I strongly recommend that careful attention be given to any variables of postulated biological significance but which must be eliminated on statistical grounds. Various statistical procedures may provide insight into the behavior of these variables, e.g., nonlinear regression, ridge regression, discriminant analysis, or nonparametric techniques.

In summary, my application of canonical correlation does not provide much insight into the

Table 4. Robust estimates of canonical correlations and variable loadings. The top portion of the table lists means and standard deviations (in parentheses) of values where one case (transect) was systematically eliminated from the analysis (n = 20). The bottom portion lists means and standard deviations when one habitat variable was eliminated (n = 9).

	Distribution variate		Habitat variate	
Rotation of Cases				
Canonical	SUMMER-FALL	0.957(0.036)	LEK	-0.827(0.051)
$\bar{r}$ = 0.950(0.011)	WINTER	0.488(0.063)	FORBS	0.826(0.047)
($\bar{P}$ = 0.0077[0.005])	SPRING	0.864(0.062)	DIV	0.625(0.043)
			PATCH	0.498(0.059)
			SPRAY	-0.811(0.048)
			COVER	0.646(0.068)
			VACOV	0.564(0.065)
			DENS	0.539(0.08)
			VADEN	0.502(0.075)
Rotation of Habitat Variables				
Canonical	SUMMER-FALL	0.969(0.024)	LEK	-0.837(0.019)
$\bar{r}$ = 0.940(0.003)	WINTER	0.506(0.065)	FORBS	0.839(0.014)
($\bar{P}$ = 0.007[0.014])	SPRING	0.846(0.069)	DIV	0.630(0.012)
			PATCH	0.497(0.012)
			SPRAY	-0.838(0.015)
			COVER	0.634(0.061)
			VACOV	0.558(0.046)
			DENS	0.527(0.073)
			VADEN	0.495(0.060)

interactions between sage grouse distribution and habitat that I could not glean from multiple regression analysis. This is not surprising since canonical correlation is somewhat of a generalization of multiple regression (Blackith and Reyment 1971). However, canonical correlation does provide a succinct and synoptic way to summarize a bulky and unwieldly data set. This is certainly one of the most important functions of multivariate statistical analysis.

ACKNOWLEDGMENTS

This work was supported by a contract from ARCO Coal Co. under the helpful supervision of James Tate, Jr. Several people have provided useful assistance in the field including B. Colenso, D. Rothenmaier, B. Holz, and J. Tate, Jr. I thank L. McDonald, K.G. Smith, and J.R. Karr for useful discussions about canonical correlation; and D.E. Capen and J. Rice for suggesting revision in the manuscript.

LITERATURE CITED

Blackith, R.E., and R.A. Reyment. 1971. Multivariate morphometrics. 412 p. Academic Press, New York, N.Y.

Boyce, M.S. 1978. Climatic variability and body size variation in the muskrats (Ondatra zibethicus) of North America. Oecologia 36:1-19.

Cohen, P., E. Gaughran, and J. Cohen. 1979. Age patterns of childbearing: a canonical analysis. Multivariate Behavioral Research 14:75-89.

Cooley, W.W., and P.R. Lohnes. 1971. Multivariate data analysis. 364 p. John Wiley and Sons, New York, N.Y.

Gauch, H.G., Jr., and T.R. Wentworth. 1976. Canonical correlation analysis as an ordination technique. Vegetatio 33:17-22.

Harris, R.J. 1975. A primer of multivariate statistics. 332 p. Academic Press, New York, N.Y.

Keammerer, W.R., and D.H. Keammerer. 1975. Floristic studies and vegetation mapping of the Coal Creek lease area. 55 p. Stoecker-Keammerer and Assoc., Boulder, Colo.

Patterson, R.L. 1952. The sage grouse in Wyoming. 341 p. Sage Books, Denver, Colo.

Seber, G.A.F. 1973. The estimation of animal abundance and related parameters. 506 p. Griffin, London.

Smith, K.G. 1981. Canonical correlation analysis and its use in wildlife habitat studies. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Thorndike, R.M. 1978. Correlational procedures for research. 340 p. Halsted Press, New York, N.Y.

ECOLOGICAL RELATIONSHIPS OF GRASSLAND BIRDS TO HABITAT AND FOOD SUPPLY IN EAST AFRICA[1]

L. Joseph Folse, Jr.[2]

Abstract.--Canonical correlation analysis was used to evaluate relationships between 17 species of grassland birds, structural parameters of the grassland habitat, and biomasses of 25 age-taxonomic categories of arthropods in a study of ecological relationships of grassland birds in the Serengeti National Park, Tanzania, East Africa. The bird-habitat analysis resulted in correlations that were interpretable both in terms of identifiable characteristics of the habitat and in terms of identifying specific groups of birds which exploited the habitat in similar ways. The bird-arthropod analysis produced correlations which were not easily interpretable relative to arthropod characteristics nor in terms of identifying groups of birds based on their association with the arthropod resource. A canonical correlation of birds with habitat plus arthropod variables had results similar to that of the bird-habitat analysis and dissimilar to the bird-arthropod analysis. These results suggest that mechanisms affecting the guild structure of grassland birds in the Serengeti are more likely associated with habitat structure than with food supply.

Key words: Arthropods; canonical correlations; grassland birds; habitat selection; multivariate analysis; Serengeti; Tanzania.

INTRODUCTION

The use of multivariate statistics in evaluating ecological relationships of birds has expanded considerably in recent years. One problem of importance that is often approached with multivariate methods is that of evaluating the structure of bird communities relative to their exploitation of habitat and/or food resources. The multivariate technique which is most suited to exploring relationships between two such groups of variables, e.g., community composition and niche parameters, is that of canonical correlation (Gittens 1979, Kshirsagar 1978, Morrison 1976). However, this technique seems to have been used little in ornithology, perhaps due to lack of adequate data from field studies.

This study illustrates an attempt to use canonical correlation analysis to identify and evaluate some relationships between grassland birds and their habitat and food supply on the Serengeti plains, Tanzania. At the outset of the study, little was known about the community composition or densities of the birds, nor about their habitat relations. Consequently, this analysis was designed to treat the following objectives: 1) to determine if there is a structural organization of the avifauna that is related to characteristics of the habitat and

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt. 05401.

[2]Assistant Professor, Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX 77843.

arthropod fauna, 2) to define recognizable groups of bird species whose members are similarly associated with habitat and arthropod characteristics, and 3) to identify specific characteristics of the habitat and arthropod fauna which may potentially affect this avifaunal structure. Objective 1 can be met by determining if there are significant canonical correlations between the birds and the environmental parameters. If groups of bird species are recognizable based on their correlations with significant canonical variates, then these may be considered as guilds due to the similarity of these species in exploiting resources, and this would satisfy Objective 2. Objective 3 may be met if the canonical variates are interpretable in terms of their associations with the habitat and arthropod variables of the analysis.

STUDY AREAS

The field study took place from May 1975 to April 1976 in the Serengeti National Park, Tanzania. Five 1-km² study sites were established to represent a range of grassland vegetation conditions in the plains portion of the park (fig. 1). Detailed descriptions of each site are available in Folse (1978). Precipitation in the Serengeti was seasonal and highly variable, both in time and space; there was an increasing gradient in precipitation from the southeast to the northwest (with 500 mm to 700 mm per anum on the plains). This variation in precipitation, coupled with intensive seasonal grazing by wildebeests, zebra, gazelle, and others, and dry season fires, resulted in a great degree of variation in vegetation structure among the different sites. Within each of the sites, however, the structure of the vegetation was reasonably uniform with major variation taking place through time (seasonally).

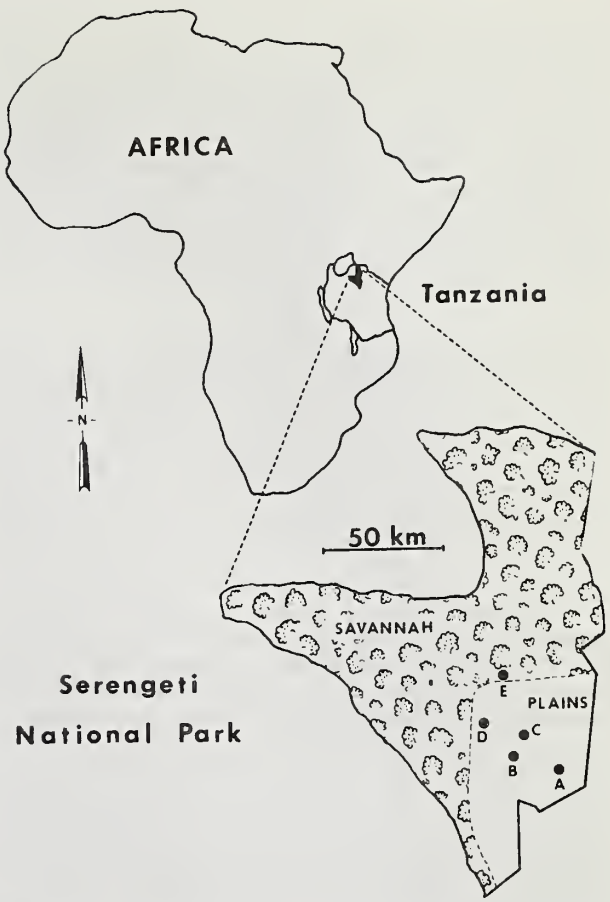


Figure 1. Location of the study sites (A through E) in the plains portion of the Serengeti National Park, Tanzania.

METHODS

The data used in these analyses were based on concurrent samples of numbers of birds of each species in the community, structural parameters of the vegetation, and numbers of arthropods in several age-taxonomic classes. Details of the sampling procedures are described in Folse (1978). These samples were taken at monthly intervals at each of the study sites (two samples were missing resulting in 58 observations). Bird numbers were sampled at each site with nine systematic transects (in groups of three). Each transect was 1 km long and 20 m wide; they were run in a vehicle and the number and identity of each bird species flushed from the transect was recorded. The sampling intensity was 18% of each study area. The bird data were converted to biomass estimates based on sampled weights and were then log-transformed prior to analysis; preliminary analyses indicated that mean values and variances of untransformed biomasses (of all biomass variables in the study) were proportional, suggesting the need for such a variance stabilizing transformation prior to analysis.

Arthropod data were collected on four sweep-net transects placed systematically on each study site. Fifty sweeps were made on each transect. Size distributions of 25 age-taxonomic categories of arthropods were available for each transect, and these were converted to biomass estimates based on sampled length-weight relationships for each group. Arthropod biomass data were also log-transformed prior to analysis.

Habitat data were collected from four transects associated with the arthropod samples. Six sample units at 20-pace intervals were used for each transect. At each sample unit, four sample points were placed in a square with 2 m diagonals. At each sample point, a thin wire was placed vertically, and the number of contacts of vegetation (both total and green) with the wire was recorded in 10 cm intervals above the ground. In addition, maximum height of emergent vegetation within a 1 m radius of the center of the sample unit was recorded. From these samples, I calculated indices of total biomass, green biomass, percent cover, percent vegetation below 10 cm (an index of vertical vegetation structure),

and vegetation height. Two additional variables were created by using values of total and green biomass lagged by one month; these variables were included to evaluate possible delayed response of birds to changes in the vegetation biomass variables. Accumulated precipitation (monthly intervals) data were available from storage gauges near each site. All biomass data were log-transformed and all the percentage data were transformed with the arcsin square root transformation (Sokal and Rohlf 1969).

Canonical correlations were run on appropriate subsamples of the data set (with monthly mean values) with the CANCORR procedure of SAS (Barr et al. 1979). I shall refer to a linear compound of original variables produced by the procedure CANCORR as a canonical variate, and coefficients of the variates associated with each original variable as factors. In two-group canonical analysis, pairs of canonical variates are produced. Each variate pair contains one variate of bird factors and one variate of environmental (habitat and/or arthropod) factors. The first pair has a maximum possible correlation. The second pair is maximally correlated, given that each is orthogonal to its corresponding variate in the first pair, and so on. Usually, only the "significant" correlations and their corresponding variate pairs are considered in evaluation of the relationships. In the usual approach to canonical analysis, the observations are plotted in variate space using either set of variates, and these observations are then clustered to look for relationships among the observations in the variate space. Thus variates are used to establish relationships among observations, and canonical correlations are used as a measure of how well the two sets of variates provide the same set of relationships among observations (site-time combinations in the present case). Interpretation of what each variate "means" is usually accomplished by considering simple correlations of each of the variables with the variate.

In the present study, emphasis was on the bird community rather than on the observations (site-time combinations); thus I wished to use canonical correlation as a technique of classification or ordination based directly on relationships of the bird species to environmental variables rather than indirectly through "association" or "distance" relationships among the birds themselves. Consequently, rather than group observations based on variate factors (as is usually done), I sought to group species of birds based on their simple correlations with the variates, using the variates as a means of establishing structural relationships between birds and environmental variables (fig. 2).

In addition to the analyses described above, I tested the homoscedasticity of the variance-covariance matrices with Wilk's generalized likelihood ratio test described by Morrison (1976:250, eq. 4). A canonical correlation also was run with all data with the variables standardized.

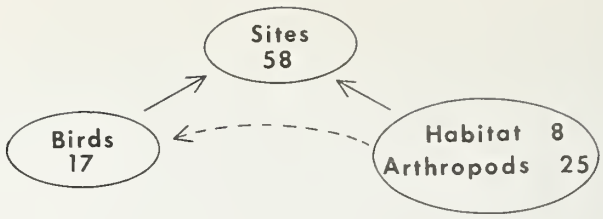


Figure 2. Relationships among bird species, environmental variables, and site-time (sites) structure of the data set. Environmental variables potentially affect composition of the bird fauna via conditions at the sites. Numbers indicate number of variables for birds and environmental characteristics and number of observations for sites.

RESULTS

Seventeen species of grassland birds were considered as study species (table 1). These were fairly small birds which depended on the grassland habitat for most of their requirements and were not dependent on specialized microhabitats such as trees, waterholes, etc. The energy requirements of these birds were satisfied mainly by arthropods (estimated 81%, Folse 1978), thus I considered arthropods as the primary food resource of the birds used in this analysis (although one species is almost wholly granivorous).

The first canonical analysis involved 17 species of birds versus all environmental variables (8 habitat and 25 arthropod). Six "significant" canonical correlations resulted with bird variates that explained 57% of the variance in the bird portion of the data set; 45% of that variance was explained by the first three correlations (table 2). Simple correlations of each bird variable with bird variates were used to determine the coordinates of each species in the correlation space associatied with the first three bird variates (fig. 3). This produced six recognizable groups of birds, each consisting of species which were associated with the environmental variables in similar ways (fig. 3).

Interpretation of the environmental variates produced by the first canonical analysis may be accomplished by examining correlations of environmental variables with environmental variates (table 3). The first variate had negative associations with high vegetation biomass, high vegetation cover, tall vegetation, and with vertical vegetation structure (the latter indicated by the positive association with percent vegetation below 10 cm). It also had negative correlations with shorthorned grasshopppers(ACRA, ACRN), spiders (ARAU), crickets (GRYU), bugs (HEMA), and caterpillars (LEPN). This variate may be interpreted as a gradient between short grass, low vegetation biomass versus tall grass, high vegetation biomass with corresponding low versus

Table 1. Common names (scientific name) of bird species considered in this study.

1.	Crowned lapwing	Vanellus coronatus
2.	Caspian plover	Charadrius asiaticus
3.	Two-banded courser	Cusorius africanus
4.	Northern white-tailed bush-lark	Mirafra albicauda
5.	Rufous-naped lark	Mirafra africana
6.	Flappet lark	Mirafra rufocinnamomea
7.	Fawn-colored lark	Mirafra africanoides
8.	Red-capped lark	Calandrella cinerea
9.	Fisher's sparrow-lark	Eremopterix leucopareia
10.	Short-tailed lark	Pseudalaemon fremantlii
11.	Richard's pipit	Anthus novaeseelandiae
12.	Sandy plain-backed pipit	Anthus vaalensis
13.	Yellow-throated longclaw	Macronyx croceus
14.	Rosy-breasted longclaw	Macronyx ameliae
15.	Capped wheatear	Oenanthe pileata
16.	Rattling cisticola	Cisticola chiniana
17.	Wing-snapping cisticola Zitting Cisticola	Cisticola ayresii Cisticola juncidis[1]

[1]These two species had broadly overlapping ranges and were indistinguishable in flight during the sample censuses.

high biomass gradient of these arthropods. The second variate was associated with low cover, low vegetation biomass (lagged) versus high cover, high vegetation biomass which was independent of vegetation height. It had no strong associations with arthropods. The third variate had a weak positive association with grasshoppers (ACRA, ACRN, TTRU) but no strong association with any of the other environmental variables. Precipitation was unimportant relative to these variates and green vegetation biomass was weakly associated with the first variate.

The second canonical analysis involved birds versus eight habitat variables. This resulted in three "significant" correlations in which the bird variates explained 39% of the variance in the bird portion of the data set (table 2). A plot of bird correlations with bird variates resulted in a group structure similar to that of the first

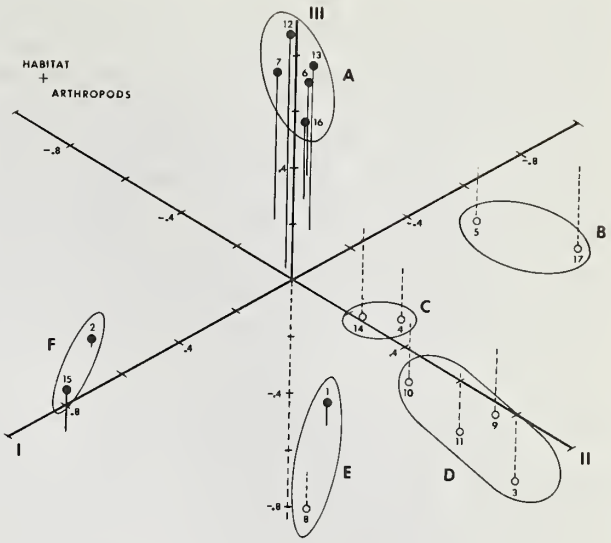


Figure 3. Bird species group structure resulting from simple correlations of bird species with the first three bird canonical variates from the complete canonical analysis (with all environmental variables included). Points below the horizontal plane of the three-dimensional plot have dashed lines and open circles. Groups indicated are potential guilds whose members are associated with environmental characteristics in similar ways. Numbers correspond to those in table 1. Group A is the "woodland" group; B the "tall-grass" group; C the "clumped-grass" group; D the "low-cover" group; E the "plains" group; and F the "short-grass-plains" group.

analysis with all environmental variables (fig. 4). Bird species 12, the sandy plain-backed pipit, was less strongly associated with the "woodland" group and the "short-grass-plains" group (species 2, the Caspian plover, and species 15, the capped wheatear) was less distinctly defined. Interpretation of environmental variates was similar to that of the complete analysis (without arthropods).

The third canonical analysis involved birds versus arthropod variables (habitat variables excluded). This resulted in four "significant" correlations in which bird variates explained 37% of variance in the bird portion of the data set (31% for the first three variates, table 2). A plot of bird correlations with bird variates (fig. 5) resulted in a completely different organization of bird species with a much less definite group structure. I can recognize 6 groups (fig. 5), only one of which occurred in the complete analysis (the "tall-grass" group--species 5, the rufous-naped lark, and species 17, the wing-snapping/zitting cistiocolas).

Canonical correlation of the full model with standardized variables resulted in different canonical variates, but identical canonical

Table 2. Canonical correlations and percent variance[1] of bird biomasses accounted for by each bird canonical variate.

	Birds vs. Habitat + Arthropods		Birds vs. Habitat		Birds vs. Arthropods	
Canonical variate	Canonical correlation	Variance %	Canonical correlation	Variance %	Canonical correlation	Variance %
1	0.999	19.1	0.984	19.6	0.987	17.9
2	0.995	14.1	0.904	13.4	0.969	5.9
3	0.989	11.9	0.802	6.4	0.954	7.3
4	0.972	3.2			0.939	5.6
5	0.963	4.6				
6	0.950	4.5				
		57.4%		39.4%		36.7%

[1]Percent variance = $(1/k)(\Sigma_j r_{ij}^2)$, where r_{ij} is the correlation of the j^{th} species with the i^{th} bird canonical variate and k is the number of bird species (17 in this case).

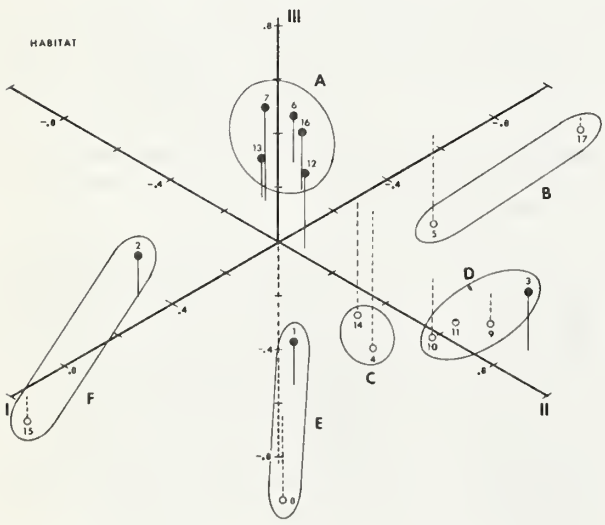


Figure 4. Simple correlations of birds species with the first three bird canonical variates with habitat variates only. See figure 3 for a detailed description. Note that group structure is similar to that of figure 3.

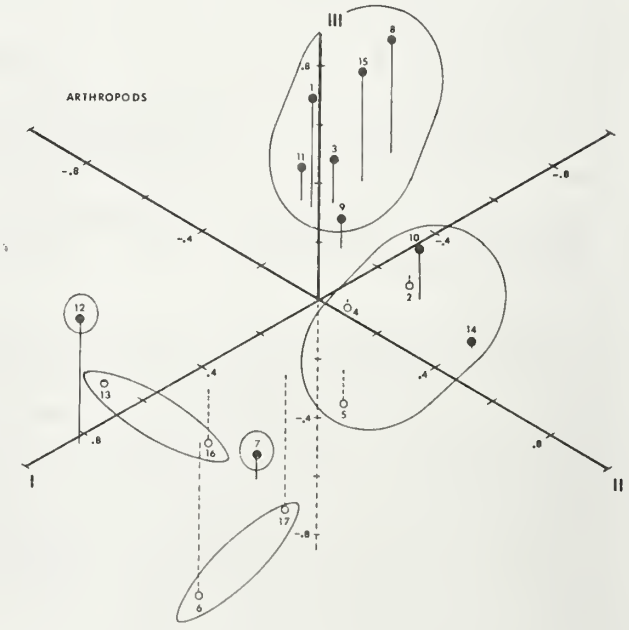


Figure 5. Simple correlations of bird species with the first three bird canonical variates with anthropod variables only. See figure 3 for a detailed description. Note that group structure is much different from that of figures 3 and 4.

Table 3. Simple correlations of environmental variables with the first three environmental canonical variates.

	Canonical variate I	II	III
Habitat:			
Total vegetation	-0.72	-0.43	-0.09
Green vegetation	-0.41	0.17	0.09
Percent cover	-0.51	-0.58	0.00
Percent veg.< 10cm	0.89	-0.04	0.20
Height	-0.89	-0.21	-0.23
Precipitation	-0.15	0.07	0.23
Tot. veg. (lagged)	-0.69	-0.50	-0.18
Gr. veg. (lagged)	-0.41	0.10	-0.12
Anthropods[1]:			
ACRA	-0.53	-0.38	0.41
ACRN	-0.52	-0.32	0.46
ARAU	-0.63	-0.31	-0.08
BLAA	-0.22	0.19	-0.16
BLAN	-0.29	0.15	-0.03
COLA	-0.37	-0.15	0.19
DIPA	-0.47	-0.17	0.13
FORA	-0.17	0.16	-0.15
GRYU	-0.67	-0.20	-0.11
HEMA	-0.52	-0.17	0.08
HEMN	-0.32	-0.25	-0.06
HOMA	-0.56	0.21	-0.03
HOMN	-0.34	-0.09	-0.11
HYMA	-0.28	0.03	0.03
ISOA	-0.14	0.17	-0.17
LEPA	-0.11	0.02	0.07
LEPN	-0.55	-0.23	0.09
MANA	-0.08	0.22	-0.10
MANN	-0.47	-0.22	0.03
NEUA	-0.08	-0.32	-0.19
OTHU	-0.05	0.08	-0.08
PHSU	-0.36	-0.02	-0.21
TETA	-0.27	-0.14	-0.02
TETN	-0.34	-0.12	0.03
TTRU	-0.33	0.04	0.63

[1]First three letters of each name symbol correspond to arthropod order or family name; last letter refers to age category: A = alate (winged adult), N = nonalate, and U = unknown (both A and N). Thus ACRA stands for Acrididae alate. OTHU is "other".

correlations and identical simple correlations of the standardized variables (as the original variables) with their corresponding variates (as expected). Results of tests of homoscedasticity of variance-covariance matrices showed that all matrices were not homoscedastic.

DISCUSSION

The question of whether or not there is a structural organization to the avifauna related to characteristics of habitat and arthropod fauna (Objective 1) can be answered affirmatively; results of all the canonical correlation analyses demonstrated it. The bird with habitat and the bird with habitat-arthropod analyses produced the same recognizable groups of bird species (Objective 2). The bird with arthropod analysis, however, produced a different organization of bird species with no clear group structure. In the complete analysis, the group structure was due primarily to the habitat characteristics rather than the arthropod characteristics, since removing arthropods from the analysis produced very little change while removing habitat characteristics led to great change in the resulting structure.

The complete analysis led to six groups of birds. The "woodland" group consisted of six species which occurred primarily at site E on the woodland-plains border. The "low-cover" group had four species which occurred throughout the Serengeti plains but were usualy in areas with extensive open ground and little ground-level obstruction. Tall grass could occur in these areas, but it was usually sparsely distributed. The "tall-grass" group (two species) occurred throughout the plains but always in association with tall grass (they were rare at site A). The greater the vertical vegetation structure and vegetation density, the greater the density of these species. The "clumped-grass" group (two species) occurred mostly at sites C and D, and were always associated with fairly dense clumps of vegetation. The "short-grass-plains" group (two species) occurred mainly at site A with little or no vertical vegetation structure, but with reasonably high percent cover. The "plains" group (two species) had a broad range throughout the Serengeti plains, but were usually associated with areas of short vegetation, low cover, and fairly reduced vegetation biomass. Three of these groups (woodland, clumped-grass, and short-grass-plains) were fairly restricted in the range of sites they used while the remaining groups (low-cover, tall-grass, and plains) each had quite broad ranges with respect to sites. However, each used specific types of microhabitat within sites and/or used a site at times of the year when habitat conditions were suitable. Each of these six groups is a good candidate for a guild (a group of species which exploit their environmental resources in similar ways). Objective 2 seems to be well satisfied.

Objective 3 has been partially satisfied by the fact that habitat variables and not arthropod variables are most important in the organization of the avifaunal group structure. Consideration of simple correlations of habitat and arthropod variables with the corresponding canonical variates suggested that the variables of most importance were vegetaion biomass, percent cover, height, and vertical structure of the vegetation. Green vegetation biomass and lagged values of vegetation biomass were also of importance. The importance of the lagged values suggests that there is some degree of inertia in the response of the birds to changes in habitat conditions.

Precipitation, per se, was unimportant. The most important arthropod variables, although minor, were grasshoppers, spiders, crickets, bugs and caterpillars.

The canonical correlation analyses of these data seemed to be very satisfactory relative to the original objectives, but were they appropriate? I had a total of 50 variables (worst case) and only 58 observations. While the number of observations was sufficient to allow the analysis to be completed technically, it was "data-poor." Thus, as an estimation technique for population parameters, the analysis would be weak but to establish relationships within my data set, the sample size was adequate, provided that "significance" of the resulting canonical correlations is regarded with caution. Since this was an exploratory analysis to look for potential relationships among bird species for further analysis and study, the technique seemed useful.

Williams (1981) pointed out that heteroscedasticity of the variance-covariance matrix used in canonical correlation analysis can lead to biased estimates of factors (coefficients of the canonical variates) and to different relationships of observations when they are plotted in canonical space versus observation space. However, these problems are much less important when considering group structure of one variable set based on simple correlations with resulting variates. This is evident from the fact that both unstandardized and standardized variables resulted in the same correlation structure whereas the corresponding canonical variates were different. However, standardizing normalized variables individually does not necessarily produce homoscedasticity since no consideration is taken of the covariance structure of the data. Although the variance-covariance matrices used in these analyses were heteroscedastic, the fact that most variables were transformed with traditionally normalizing transformations, and the correlations rather than variate factors were used in establishing avian groups, should have led to a fairly stable analysis. Dropping out half of the variables for the habitat analysis (the arthropod variables) resulted in little change in the avifaunal group structure, which suggests that heteroscedasticity characteristics of the arthropod portion of the variance-covariance matrix had little effect in biasing the results.

ACKNOWLEDGMENTS

I thank William B. Smith for discussions on methodology, the Caesar Kleberg Program in Wildlife Ecology for sponsoring my research in Africa, and T. Mcharo and the Director and trustees of Tanzania National Parks for permission to live and work in the Serengeti National Park. This is contribution number TA16128 of the Texas Agricultural Experiment Station.

LITERATURE CITED

Barr, A.J., J.H. Goodnight, J.P. Sall, W.H. Blair, and D.M. Chilko. 1979. SAS User's Guide. 494 p. SAS Institute, Raleigh, N. C.

Folse, L.J., Jr. 1978. Avifauna-resource relationships on the Serengeti Plains. Ph.D. Dissertation. 107 p. Texas A&M University, College Station, Tex.

Gittens, R. 1979. Ecological applications of canonical analysis. p. 309-535. In Orloci, L., C.R. Rao, and W.M. Stiteler, editors. Multivariate methods in ecological work. 550 p. International Co-operative Publishing House, Fairland, Md.

Kshirsagar, A.M. 1978. Multivariate analysis. 534 p. Dekker, New York, N.Y.

Morrison, D.F. 1976. Multivariate statistical methods. Second edition. 415 p. McGraw-Hill, New York, N.Y.

Sokal, R. R., and F.J. Rohlf. 1969. Biometry. 776 p. Freeman, San Francisco.

Williams, B.K. 1981. Discriminant analysis in wildife research: theory and applications. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

DISCUSSION

NOVA SILVY: Could arthropods be correlated with the vegetation structure in a manner such that they contributed nothing new to the analysis when they were added into the data sets?

L. JOSEPH FOLSE, JR.: The arthropod distribution is very much determined by the vegetation distribution; however, the relationships do not seem to be concordant with the bird-habitat relationships. If they were, the bird-arthropod canonical correlation would not result in a completely different organizational structure.

HABITAT ASSOCIATIONS OF BIRDS BREEDING IN CLEARCUT DECIDUOUS FORESTS IN WEST VIRGINIA[1]

Brian A. Maurer[2], Laurence B. McArthur[3], and Robert C. Whitmore[4]

Abstract.--Associations between vegetation structure and 34 bird species in four forested areas of various stages of clearcut regrowth were examined using principal components analysis. Relative frequencies for each bird species were determined during three breeding seasons and used to weight habitat variables. The resulting data matrix (34 species x 8 habitat variables) was subjected to principal components analysis using a standardized covariance matrix.

The first principal component was negatively correlated with percent and mean height of low vegetation, and positively correlated with percent litter and the number, height, and percent of canopy layers. The first principal component separated early successional species from late successional species. The second principal component was positively correlated with percent slash and the number of trees less than 12.7 cm dbh. This component separated mid-successional species from earlier and later successional species. The first two components explained 90% of the variation and thus seemed to be an adequate description of the habitat associations of most species. The third component, however, was useful in separating a few of the mid-successional species, with species that foraged mainly on the ground having higher values than species that foraged on small trees and shrubs. The third component was positively correlated with percent litter and negatively correlated with the number of trees less than 12.7 cm dbh. Field methods used in this study appear to be most applicable where it is impractical to use more conventional methods of collecting habitat association data, e.g., territory mapping, or use of male singing perches.

Key words: Clearcut; deciduous forests; habitat associations; habitat ordination; nongame management; passerine birds; principal components analysis; West Virginia.

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Graduate Research Assistant, Division of Forestry, West Virginia University, Morgantown, WV 26506. Present address: Center for Quantitative Studies, University of Arizona, Tucson, AZ 85721.

[3]Wildlife Biologist, Division of Natural Resources, P. O. Box 67, Elkins, WV 26241.

[4]Associate Professor, Division of Forestry, West Virginia University, Morgantown, WV 26506.

INTRODUCTION

Current field techniques for assessing habitat relationships are based on measuring habitat variables in activity centers of individual organisms. In the case of birds, male singing perches are often used to determine activity centers. This paper presents another method of obtaining habitat association data that can be used to compare overall habitat preferences of bird species breeding in several different habitats. The procedure analyzes variation between species by obtaining a mean value for each species for each of several habitat variables, and subjecting the resulting data matrix to a principal components analysis.

STUDY AREAS

Data on bird populations and habitat structure were collected on four watersheds in the Fernow Experimental Forest, 4.8 km southeast of Parsons, Tucker Co., West Virginia. The watersheds have been used by the USDA Forest Service to assess impacts of various logging and herbicide treatments on forest hydrology. Logging and herbicide treatments have created several different habitats which attract different bird communities.

Watershed 4 (WS4), a 39-ha area used as a control in the hydrology experiments, was logged in 1911, and has remained relatively untouched since then. Trees killed by chestnut blight were salvaged during the early 1940's. Tree species common in this area included sugar maple, red maple, red oak, chestnut oak, black birch, black cherry, American beech, and yellow poplar (scientific names of plants and birds are given in appendix I).

Watershed 3 (WS3), 34 ha in size, was clearcut between July 1969 and May 1970, except for a strip of trees along the main drainage. Remaining trees were removed during the winter of 1972-1973, and the area has revegetated naturally since then. Tree species common in this area included saplings of many mature forest species, as well as sassafras and pin cherry.

Watershed 7 (WS7), a 24-ha watershed, was partially clearcut in 1964, and regrowth was suppressed with herbicide treatments until 1967, when the rest of the watershed was clearcut. The area was kept barren for another 2 years before allowing regrowth to resume. Saplings of several mature forest trees were common in this area. Sassafras and pin cherry were also present, however, staghorn sumac was common in the watershed.

Watershed 6 (WS6), 22 ha, was treated with herbicides in a manner similar to WS 7. The watershed has been further treated with herbicides several times since 1971 to discourage natural hardwood and herbaceous revegetation, and to encourage establishment of Norway spruce. This area had a dominant ground cover of ferns, with several species of saplings and blackberry occurring sporadically throughout the watershed.

Table 1. Three-year means for vegetation structure variables in each watershed.

Variable	WS4	WS3	WS7	WS6
Litter (%)	87.3	83.6	68.7	59.7
Slash (%)	7.6	17.2	11.7	11.3
Herbaceous vegetation (%)	42.2	24.7	78.9	85.7
$\bar{X}$ height herb. veg. (cm)	27.9	29.7	61.7	59.3
Canopy layers (no.)	2.1	1.4	1.1	0.2
Max. canopy height (m)	21.3	6.7	4.3	2.1
Canopy cover (%)	96.8	90.9	77.8	4.9
Trees <12.7 cm (no.)	0.6	2.3	1.1	0.2

METHODS

Avian populations were censused during the breeding seasons of 1977, 1978, and 1979 using 30-m belt transects established in each area. All areas were visited about the same number of times each year. Transects were walked between sunrise and 0730 EDT at a predetermined pace, and the number of singing males of each bird species was recorded.

On 1m x 1m plots located randomly within the belt transects, 13 vegetation characteristics were measured in each watershed during July. In statistical analyses, only eight variables (table 1) were actually used to minimize duplication of information.

We assumed that vegetation characteristics prevalent in a watershed during a given year directly or indirectly influenced the relative abundance of bird species in that watershed. Making this assumption, we weighted the mean value of each habitat variable by relative frequency of a bird species in a watershed for each year. Years in which a species was more common in a watershed were weighted more than years in which a species was relatively less common. This was accomplished by dividing relative frequency of a species for a given watershed-year combination by the sum of all relative frequencies for that species over the 12 watershed-year combinations. This procedure is illustrated for the red-eyed vireo in table 2.

Table 2. Relative frequency of red-eyed vireos (RF_i) and mean % canopy cover ($\bar{X}_i$) for 3 years in four watersheds. The procedure of obtaining weighted mean values ($\bar{X}_w$) used in principal components analysis is illustrated by these data.

Year	Watershed	RF_i	$\bar{X}_i$
1977	4	0.32	92.7
	3	0.06	84.6
	7	0.02	66.0
	6	0	6.0
1978	4	0.49	98.7
	3	0.23	94.2
	7	0.15	84.2
	6	0.02	2.5
1979	4	0.23	99.4
	3	0.28	95.6
	7	0.19	88.0
	6	0.04	5.0

$$\Sigma\ RF_i = 2.03$$

$$\bar{X}_w = \frac{\sum_{i=1}^{12} RF_i \bar{X}_i}{2.03} = 91.3\%$$

Table 3. Results of principal components analysis using weighted averages of 8 habitat variables for 34 bird species.

Component	1	2	3
Variation explained	67.46%	23.47%	5.50%
Cumulative variation	67.46%	90.93%	96.43%
Variable	Correlations with original variables		
Litter	0.85**	0.23	0.42**
Slash	-0.38*	0.86**	0.32
Herb. veg.	-0.93**	-0.29	0.07
$\bar{X}$ ht. herb. veg.	-0.96**	-0.01	-0.04
Canopy layers	0.98**	-0.14	0.01
Max. canopy ht.	0.90**	-0.41*	0.08
Canopy cover	0.95**	0.11	-0.20
Trees < 12.7 cm	0.27	0.90**	-0.33*

* $P < 0.05$, ** $P < 0.01$

Entries in the data matrix (34 species x 8 habitat variables) to be used for a principal components analysis were weighted mean values for each habitat variable for each species. Calculations described above were done using the PROC MATRIX procedure of Statistical Analysis System (SAS, Helwig and Council 1979). A correlation matrix for the habitat variables was calculated and a principal components analysis (PCA) was done on PROC MATRIX using a program written by the second author. Output included eight eigenvalues and eigenvectors, correlations of original variables with each principal component, scores for each bird species on the principal components, and a plot of species scores on the components. Copies of the PCA program are available from the authors.

RESULTS

Means for all habitat variables are given in table 1. Inspection of these data were useful in obtaining an overall impression of the structural characteristics of each watershed.

Results of the PCA (table 3) indicated that the first three principal components accounted for most variation between bird species in their associations with habitat variables. Correlations of original variables with the first three components demonstrated that variables which tended to have high values in WS4 (mature forested watershed) were positively correlated with the first component, and variables with low values for WS4 were negatively correlated with the first component (table 3). The second principal component was significantly correlated with variables that either had high values or low values for WS3, a mid-successional area.

A plot of species scores on the first two principal axes indicated that species were separated into three distinct groups (fig. 1). The first group had low scores on the first two components and was representative of species which preferred early-successional or open habitats. Some of these species, such as song sparrows and prairie warblers, were exclusively limited to WS6 and WS7 (McArthur 1980). The second group had high scores for the first principal component, and low scores on the second principal component. The group was composed of species which were primarily mature forest species. The third group had high values for both components and included species typical of mid-successional habitats.

The third principal component explained about 6% of the variation. This component separated some of the early to mid-successional birds from each other. Species which forage primarily from trees and shrubs, such as Canada warblers (Bent

1953) had negative scores on this axis, while ground foraging birds, such as brown thrashers and rufous-sided towhees (Bent 1968), had positive scores. The third principal component was postively correlated with litter and negatively correlated with number of small trees.

DISCUSSION

Previous methods of collecting habitat association data for birds have been to locate an activity center of an individual and measure habitat variables at that spot. In several studies, habitat variables have been measured on small plots (0.05 ha) centered around perches where singing males were observed (James 1971; Whitmore 1975, 1977; Smith 1977). However, the physical structure of some habitats restricts the ability of observers to approach perches of singing males, and could introduce a bias during data collection. For example, in eastern deciduous forests, early stages of regrowth produce dense stands of saplings laced with greenbriar and other deterrents to movement. WS3 in the present study was such an area. Though we could easily hear singing males, the problems in setting up even a small plot were formidable. Another method of obtaining habitat association data has been to locate territories (McArthur 1980, Rice 1978) or nests (Wray and Whitmore 1979) and measure vegetation in these areas. Again, in some habitats the physical structure of vegetation may impede efforts to collect data. The method we have presented alleviates many logistical difficulties that might be encounterd while collecting habitat association data in areas such as WS3.

Limitations of our analysis are as follows. First, the resolution obtained may not be as fine as might be desired in some situations. From a theoretical viewpoint, small differences in habitat preference between ecologically similar species may not be distinguishable. Also, though the method is suggestive, it is not strictly predictive. In addition, the problem of measuring the appropriate variables is always present. This problem can be partially alleviated by using literature to obtain ideas as to which habitat variables might be important. Finally, this method does not analyze within-species variation.

This method of assessing habitat associations should be useful in habitat management. Specific information is summarized by such an analysis on types of habitat characteristics a given bird species is associated with. Though this does not necessarily imply a cause-effect relationship between habitat variables and bird species abundanace, information provided by the PCA can give at least a rough estimate of how species will respond to alterations of habitat due to land use practices.

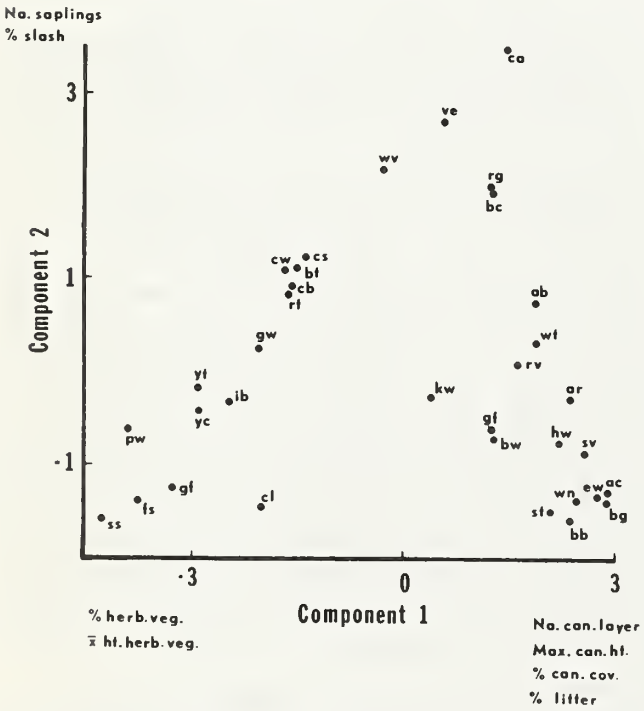


Figure 1. Ordination of 34 bird species on the first two principal components obtained from a PCA of 8 habitat variables. Species codes given in appendix.

ACKNOWLEDGMENTS

We wish to thank B. Griffin, J. Hamilton, S. Harman, R. Kidwell, R. Manna, B. Parker, and G. Seidel for their assistance in the field. The Timber and Watershed Laboratory, USDA Forest Service in Parsons, West Virginia, kindly provided housing arrangements and served as a source of information about the study areas. E.J. Harner provided statistical advice and encouragement. This research was supported, in part, by McIntire-Stennis funds from USDA (SEA) and is approved by the Director, West Virginia Agriculture and Forestry Experiment Station as Scientific Article No. 1661.

LITERATURE CITED

Bent, A.C. 1953. Life histories of North American wood warblers. United States National Museum Bulletin 203:1-734.

Bent, A.C. 1968. Life histories of North American cardinals, grosbeaks, buntings, towhees, finches, sparrows, and allies. United States National Museum Bulletin 237:1-602.

Helwig, J.T., and K.A. Council, editors. 1979. SAS user's guide. 494 p. SAS Institute, Raliegh, N. C.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

McArthur, L.B. 1980. The impact of various forest management practices on passerine community structure. Ph.D. Dissertation. West Virginia University, Morgantown, W. Va.
Rice, J.R. 1978. Ecological relationships of two interspecifically territorial vireos. Ecology 59:526-538.
Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58:810-819.
Whitmore, R.C. 1975. Habitat ordination of passerine birds of the Virgin River Valley, southeast Utah. Wilson Bulletin 87:65-74.
Whitmore, R.C. 1977. Habitat partitioning in a community of passerine birds. Wilson Bulletin 89:253-265.
Wray, T., Jr., and R.C. Whitmore. 1979. Effects of vegetation on nesting success of vesper sparrows. Auk 96:802-805.

Appendix I

Scientific names of bird and plant species mentioned in the text. Codes for birds given in parentheses.

Birds

Great-crested flycatcher (gf)	*Myiarchus crinitus*
Acadian flycatcher (ac)	*Empidonax virescens*
Eastern wood pewee (ew)	*Contopus virens*
Black-capped chickadee (bc)	*Parus atricapillus*
White-breasted nuthatch (wn)	*Sitta carolinensis*
Gray catbird (cb)	*Dumetella carolinensis*
Brown thrasher (bt)	*Toxostoma rufrum*
Wood thrush (wt)	*Hylocichla mustelina*
Veery (ve)	*Catharus fuscescens*
Cedar waxwing (cw)	*Bombycilla cedrorum*
Solitary vireo (sv)	*Vireo solitarius*
White-eyed vireo (wv)	*V. griseus*
Red-eyed vireo (rv)	*V. olivaceus*
Black-and-white warbler (bw)	*Mniotilta varia*
Golden-winged warbler (gw)	*Vermivora chrysoptera*
Black-throated green warbler (bg)	*Dendroica virens*
Black-throated blue warbler (bb)	*D. caerulescens*
Chestnut-sided warbler (cs)	*D. pensylvanica*
Prairie warbler (pw)	*D. discolor*
Ovenbird (ob)	*Seiurus aurocapillus*
Common yellowthroat (yt)	*Geothlypis trichas*
Yellow-breasted chat (yc)	*Icteria virens*
Kentucky warbler (kw)	*Oporonis formosus*
Hooded warbler (hw)	*Wilsonia citrina*
Canada warbler (ca)	*W. canadensis*
American redstart (ar)	*Setophaga ruticilla*
Scarlet tanager (st)	*Piranga olivacea*
Cardinal (cl)	*Cardinalis cardinalis*
Rose-breasted grosbeak (rg)	*Pheucticus ludovicianus*
Indigo bunting (ib)	*Passerina cyanea*
American goldfinch (gf)	*Spinus tristus*
Rufous-sided towhee (rt)	*Pipilo erythropthalmus*
Field sparrow (fs)	*Spizella pusilla*
Song sparrow (ss)	*Melospiza melodia*

Plants

Red maple	*Acer rubrum*
Sugar maple	*A. saccharum*
Black birch	*Betula lenta*
American beech	*Fagus grandifolia*
Yellow poplar	*Liriodendron tulipfera*
Norway spruce	*Picea abies*
Pin cherry	*Prunus pensylvanica*
Black cherry	*P. serotina*
Chestnut oak	*Quercus prinus*
Red oak	*Q. rubra*
Staghorn sumac	*Rhus typhina*
Blackberry	*Rubus* spp.
Sassafras	*Sassafras albidum*
Greenbriar	*Smilax* spp.

DISCUSSION

BARRY NOON: Did your principal components analysis give you any additional insights into the species-habitat associations that were not apparent from a simple list of species by habitat type?

BRIAN MAURER: Yes, the PCA helped to identify specific relationships between bird species and habitat variables. For example, both chestnut-sided warblers and common yellowthroats were found on three areas. The PCA showed that yellowthroats were more strongly associated with variables dealing with the abundance and height of herbaceous vegetation, while chestnut-sided warblers were not. These types of relationships are not apparent by simply examining species lists. In the example just given, since both species would appear on lists for the same habitats there would be no way to further differentiate between their specific habitat requirements, other than drawing speculative conclusions from the life history literature.

KEN MORRISON: Why did you use the matrix procedure of SAS rather than PROC FACTOR with METHOD=PRINT to obtain your principal components?

BRIAN MAURER: It is easier to get the factor scores from PROC MATRIX, because you do not need to call an additional procedure. Also, the researcher can easily modify a PROC MATRIX program to fit the particular needs of the data.

PRINCIPAL COMPONENTS ANALYSIS OF DEER HARVEST—LAND USE RELATIONSHIPS IN MASSACHUSETTS[1]

Philip J. Sczerzenie[2]

Abstract.--Relationships of deer harvests with land use and forest cover types in Massachusetts' 351 townships in 1951 and 1971 were investigated using principal components (PC) techniques. PC analysis and PC regression methods are described and their value in reducing the dimensionality of the land-use and forest-type data is emphasized. Components on softwood vs. urban land, stand-age and hardwoods + farmland accounted for over 40% of the X data variability. Increasing softwood composition, decreasing age, and increased association of hardwoods and farmlands were significant positive effects when related to deer harvest levels. PC regression, under well-defined criteria for component deletion, allowed estimation of effects in the original beta-space with reduced variances on those effects while maintaining the statistical integrity of the data sets.

Key words: Forest types; land use; multiple regression; principal components; white-tailed deer.

INTRODUCTION

Massachusetts' legal deer harvests began in 1910 under supervision of the state's Division of Fisheries and Wildlife. Harvest levels increased, except during depression and war years, to a record take of 4,887 deer in 1958. Harvests declined precipitously after 1958, until in 1966 mandatory deer checking, and in 1967 anterless deer permit systems, were instituted.

Two concurrent factors appeared to have caused the decline in harvests. First, either-sex hunting from 1910 to 1966 resulted in an overharvest of does, so the reproductive portion of the herd was unable to replenish itself. Second, farm abandonment and secondary succession that led to buildup of the Massachusetts herds in the late 1800's no longer provided enough productive habitat. Instead, a decline in forage and cover was occurring due to urbanization and forest succession. The antlerless permit system was established to remedy the former problem, and while it was successful in limiting doe take and increasing buck harvests (McDonough and Pottie 1979), it had no effect on the second problem.

The present study was undertaken to quantify effects of different land uses and forest types on deer harvests so that alternative approaches to habitat management could allocate management effort optimally. The 20-year period between samples would tend to strengthen conclusions drawn on the basis of effects that remained the same.

DATA DESCRIPTION

The dependent variable, deer harvest, was the mean 5-year kill in each of 351 townships in Massachusetts from 1949-53 and 1969-73. As

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Graduate Research Assistant, Dept. Forestry and Wildlife Management, Holdsworth Hall, University of Massachusetts, Amherst, MA 01003. Present address: Ketron, Inc., 18th Floor Rosslyn Center, 1700 North Moore Street, Arlington, VA 22209.

recommended by Labisky et al. (1964), kill was transformed using $\log_{10}(1 + \text{kill/mi}^2)$. Deer harvest averages centered on the years 1951 and 1971 to coincide with land-use surveys of the state based on aerial photographs (MacConnell 1975). Aerial photos were land-use typed to approximately 3.5 acres; this information was transferred to topographic sheets; and the area of each of 104 types computed and recorded for each township.

I used 35 types as independent variables: abandoned fields (AF), pasture (PS), urban land (UL), and 32 forest types (table 1). Acreage of each type was transformed to arcsin square root percent (Steel and Torrie 1960:158). Total acreage of the 35 types was 4,463,360 in 1951 and 4,480,327 in 1971 representing 85.9% and 86.2% of the state's total surface area of 5.2 million acres in respective years.

PRINCIPAL COMPONENT ANALYSIS

Method

Principal component analysis (PCA) has been used by ecologists as one of a number of multivariate techniques to describe habitat preferences of various species of North American breeding birds (James 1971, Smith 1977) and to describe site differences in relation to vegetation communities (Austin 1968, Page 1976).

PCA is helpful because it reduces the dimensionality of a data set, such as an array of environmental variables assumed to influence the species under consideration. A more parsimonious description of habitat influence is possible, therefore, assuming the principal components have some reasonable biological interpretation.

The original data (X), expressed as a matrix of simple correlations (by standardizing $\underline{X}$ and premultiplying by $\underline{X}'$), is transformed into a matrix of principal component scores, $\underline{z}_i$, (regressors in PC regression). This procedure is accomplished by creating a set of eigenvectors, $\underline{a}_i$, and eigenvalues, λ_i.

Each $\underline{a}_i$ is a linear combination of loadings between −1 and +1 on each X_i, constructed so that, in sequence (i=1,2,...,k), the maximum possible amount of variability in the X data is accounted for, with the stipulation that each eigenvector be orthogonal with every other $\underline{a}_i$ ($\underline{a}_i' \cdot \underline{a}_j = 0$).

Each eigenvalue, λ_i, represents the portion of variation explained by its corresponding eigenvector. The following equations apply:

$$\underline{X}\,\underline{a}_i = \underline{z}_i \qquad (1)$$

$$\underline{z}'_i \cdot \underline{z}_i = \lambda_i = \underline{a}'_i\,\underline{X}'\underline{X}\,\underline{a}_i \qquad (2)$$

and the resultant matrix of orthogonal eigenvalues is

$$\begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & & & & \lambda_k \end{bmatrix}$$

When based on the correlation matrix ($\underline{X}'\underline{X}$), $\Sigma\lambda_i = k$, where k is the number of independent (X) variables.

Table 1. Classification of 32 forest types used as independent variables (MacConnell 1973).

Forest composition:			
H:	hardwoods	HS:	hardwood dominant mixed forest
S:	softwoods	SH:	softwood dominant mixed forest
Forest height classes:			
1:	one to 20 feet	2:	21 to 40 feet
3:	41 to 60 feet		
4:	61 to 80 feet	6:	uneven heights (3 or more classes present)
Crown closure classes:[1]			
A:	81 to 100% crown closure	B:	30 to 80% crown closure

[1]Height classes 1 and 6 have no crown closure associated.

Table 2. Eigenvalues and cumulative porportion of $\underline{X}$ variation of 1951 and 1971 data on 35 land uses.

		1951			
Eigenvalues					
7.98896	4.16814	3.29646	2.63986	1.66862	1.41153
1.35130	1.16983	1.06800	0.84554	0.82070	0.73461
0.69440	0.65197	0.62429	0.59965	0.52160	0.49797
0.45877	0.40060	0.36848	0.35173	0.34111	0.29181
0.27749	0.25885	0.22551	0.21778	0.19392	0.17896
0.16296	0.16006	0.14834	0.11799	0.09221	
Cumulative proportion of total variance of independent variable					
0.22826	0.34735	0.44153	0.51695	0.56463	0.60496
0.64357	0.67699	0.70751	0.73166	0.75511	0.77610
0.79594	0.81457	0.83241	0.84954	0.86444	0.87867
0.89178	0.90322	0.91375	0.92380	0.93355	0.94188
0.94981	0.95721	0.96365	0.96987	0.95741	0.98053
0.98518	0.98976	0.99399	0.99737	1.00000	
		1971			
Eigenvalues					
6.82717	4.99291	4.24402	2.84598	2.43800	1.39437
1.24683	1.07258	0.97177	0.79921	0.75289	0.65541
0.56755	0.51426	0.50080	0.46745	0.42616	0.41565
0.39114	0.36751	0.31664	0.30569	0.28716	0.28383
0.27362	0.23534	0.22413	0.20672	0.17137	0.16534
0.15844	0.14815	0.12635	0.11795	0.08863	
Cumulative portion of total variance of independent variable					
0.19503	0.33679	0.45895	0.54026	0.60992	0.64976
0.68538	0.71602	0.74379	0.76662	0.78813	0.80686
0.82308	0.83777	0.85208	0.86543	0.87761	0.88949
0.90066	0.91116	0.92021	0.92824	0.93715	0.94526
0.93507	0.95980	0.96620	0.97211	0.97700	0.98173
0.98625	0.99049	0.99410	0.99747	1.00000	

The first few principal components normally account for the bulk of $\underline{X}$ variability while the last few explain a negligible portion and can, in theory, be disregarded. This reduces the dimensionality of $\underline{X}$ to $\underline{z}_i$, less than k, and, if $\underline{z}_i$ can be given some interpretation (usually by inspection of eigenvector loadings), a theory about the underlying structure of $\underline{X}$ can be expounded. With respect to a dependent variable, the relationship between $\underline{Y}$ and each $\underline{z}_i$ may be examined, in many cases graphically, to determine how the underlying $\underline{X}$ structure affects the species of interest. For a more complete treatment of principal component analysis the reader may consult Johnston (1972) or Nichols (1977).

Results

Table 2 lists eigenvalues of the 1951 and 1971 X data (land uses and forest types) in order of importance and the cumulative proportion of total $\underline{X}$ variance explained by successive eigenvalues. In both cases, the first four eigenvalues account for more than half the variability, leaving the remaining 31 components to explain less than 50%. Figure 1 illustrates the striking similarities between eigenvectors $\underline{a}_1$ each year, eigenvectors $\underline{a}_2$ each year, and $\underline{a}_3$ in 1951 and $\underline{a}_4$ in 1971. Each pair of eigenvectors measures the same relationship in $\underline{X}$ in both years although figure 1A and 1C are mirror images due to pecularities in computation of $\underline{a}_i$.

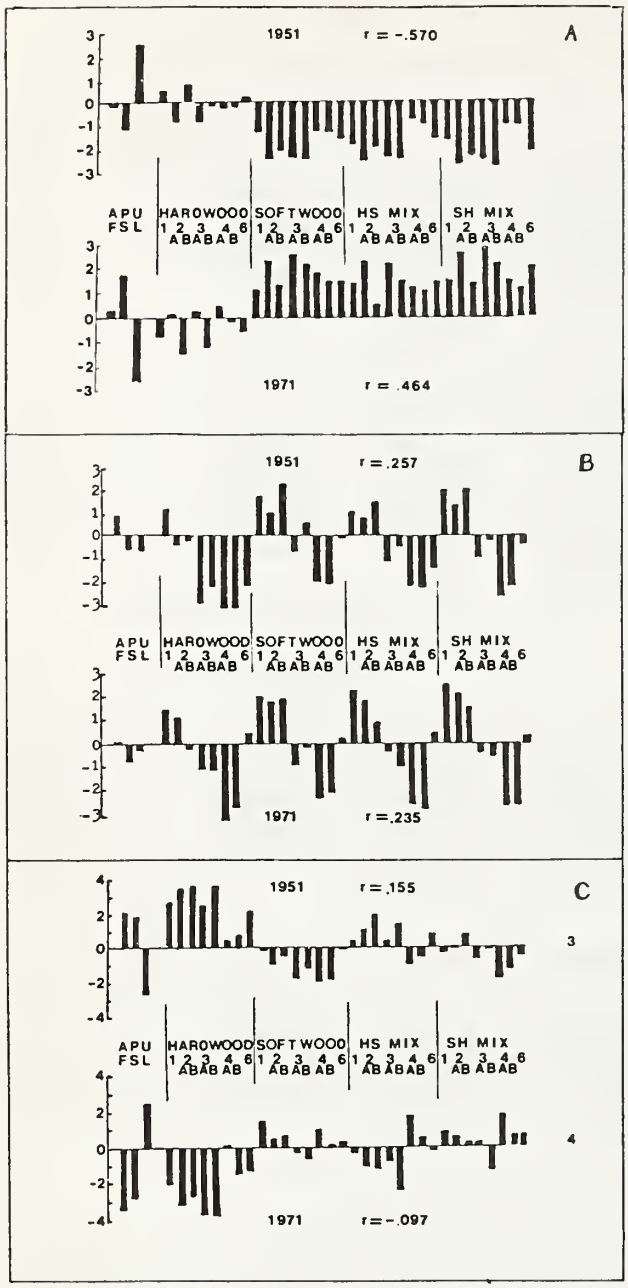


Figure 1. Eigenvectors of 35 Massachusetts land types: (A) eigenvectors $\underline{a}_1$, (B) eigenvectors $\underline{a}_2$, (C) eigenvectors $\underline{a}_3$,$\underline{a}_4$.

All forests with a softwood component have relatively heavy loadings in $\underline{a}_1$ while urban land is weighted in the opposite direction. This can be termed the softwood vs. urban land (SUL) effect. It accounts for about 20% of $\underline{X}$ variability in each year. The second eigenvector is loaded positively for young forest of all types, less positively and eventually negatively for stands of intermediate age, and negatively for older age forest stands. This effect of stand age (STA) remains the same over the 20-year sampling period. Eigenvectors $\underline{a}_3$ of 1951 and $\underline{a}_4$ of 1971 can be described as the hardwood/farmland (HF) effect since they are loaded for hardwood forest types and AF and PS in both years.

Three corresponding pairs of eigenvectors; $\underline{a}_1$, $\underline{a}_2$, and $\underline{a}_3$ of 1951 and a_1, $\underline{a}_2$, and $\underline{a}_4$ of 1971; account for 44 and 42 percent of $\underline{X}$ variability respectively, therefore, the dimensionality of land-use and forest type data has been reduced, to a great extent, to information on three independent axes of variability that have clear structures.

Eigenvectors act on $\underline{X}$ to form principal component scores, $\underline{z}_i$, as follows. Each township has a given percentage composition of the 35 land-use types that constitute, when transformed, the observations X_1, X_2,...,X_k for the township in that year. Each X_i is multiplied by the corresponding eigenvector element and the products summed to form z_i for the township according to equation (1). If a township has, for example, large percentages of older hardwoods and urban land in 1971, its score for z_1 would be negative, for z_2 negative and for z_4 positive.

Relationship with Deer Harvest

Relationships between principal axes of land-use and forest-type variation and corresponding township deer kill levels are summarized in table 3. Component 1 in each year has the highest eigenvalue, explains the highest percentage of $\underline{X}$ variation, and has the highest simple correlation with deer harvest. Because the $\underline{a}_1$'s are mirror images their corresponding r's are of opposite sign. One can conclude townships with low amounts of urban land and higher acreage of softwoods are more productive of deer. Component 2 in each year explains less $\underline{X}$ variability and has a lower, positive correlation with deer harvest, thus, stand-age is an important determinant of a township's productivity for deer. Those with older stands, in general, have fewer deer harvested.

The third PC effect, hardwood/farmland, has a relatively low correlation with deer harvest and, therefore, although townships with more hardwoods generally have higher harvests, one would tend to manage for softwoods or mixed forests rather than hardwoods since the latter is a much stronger relationship and appears to contradict the former finding. The combination of hardwoods and farmland, on the other hand, may serve as an indication of the relative lack of urbanization in a township and, in this respect, would reasonably be associated with higher deer harvests.

Table 3. Comparison of three important principal components of 1951 and 1971 data on three land-use and 32 forest types.

Year	Component	Eigenvalue	$\underline{X}$ Variation explained (%)	r	b[1]
1951	1	7.989	22.8	-0.570	-0.148
1971	1	6.826	19.5	0.464	0.133
1951	2	4.168	11.9	0.257	0.092
1971	2	4.993	14.3	0.235	0.079
1951	3	3.296	9.4	0.155	0.063
1971	4	2.846	8.1	-0.097	-0.043

[1]Prinicpal component beta coefficients, $P < 0.01$.

PRINCIPAL COMPONENTS REGRESSION

Method

Least Squares Model

The preceding PC analysis was done as an intermediate stage in what many researchers consider the primary purpose of multivariate analysis, estimation of coefficients on the original variable set. Thus, the value of principal components lies in their usefulness in solving the general linear model:

$$Y_j = B_o + B_1X_{1j} + B_2X_{2j} + \ldots + B_kX_{kj} + u_j \quad (3)$$

where Y_j are observations on the dependent variable, $(j=1,2,\ldots,n)$ B_o is the intercept, B_i is the effect of the i^{th} independent variable on Y, $(i=1,2,\ldots,k)$, and u_j is a randomly distributed error or disturbance term. In matrix notation (3) becomes

$$\underline{Y} = \underline{X}\,\underline{B} + \underline{e} \quad (4)$$

where $\underline{Y}$ is an n x 1 vector of dependent variable observations, $\underline{X}$ is an n x k matrix of independent variable observations, $\underline{B}$ is a k x 1 vector of coefficients on $\underline{X}$, and $\underline{e}$ is a residual term (substituted for u) that is the difference between $\underline{Y}$ observed and $\underline{Y}$ predicted using our estimate of $\underline{B}$.

Using the matrix $\underline{A}$ of k eigenvectors derived from the matrix $\underline{X}'\underline{X}$ of correlations, we form the n x k matrix of z scores $\underline{Z}$ by the k x 1 vector $\underline{d}$ of coefficients on the scores as

$$\underline{Y} = \underline{XA}'\underline{A}\,\underline{B} + \underline{e} = \underline{XA}\underline{d} + \underline{e} = \underline{Z}\underline{d} + \underline{e} \quad (5)$$

The vector $\underline{d}$ of PC coefficients is easily solved for

$$\underline{d} = (\underline{Z}'\underline{Z})^{-1}\,\underline{Z}'\underline{Y} + (\underline{Z}'\underline{Z})^{-1}\underline{Z}'\underline{e} \quad (6)$$

since

$$\underline{Z}'\underline{Z}^{-1} = \begin{bmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & \\ & & \ddots & \\ & & & 1/\lambda_k \end{bmatrix}$$

The second term on the right in (6) is zero. This leads to the ordinary least squares (OLS) solution when $\underline{d}$ is transformed back to the original beta-space using $\underline{B} = \underline{A}\underline{d}$.

Component Deletion

The PC regression (PCR) solution, therefore, calls for deletion of one or more components to rid the data structure of "noise" that provides little in the form of information about $\underline{X}$ variation (recall cumulative proportion of variability from table 2) but adds considerably to variance about $\underline{B}$.

The data structure is partitioned as:

$$\underline{Y} = \underline{XA}_1\,\underline{d}_1 + \underline{XA}_2\underline{d}_2 + \underline{e} = \underline{Z}_1\underline{d}_1 + \underline{Z}_2\underline{d}_2 + \underline{e} \quad (7)$$

and deletion of the second portion is required. This is equivalent to the restriction $\underline{XA}_2\underline{d}_2 = 0$ (Fomby and Hill 1978). If the restriction is true, the estimator remains unbiased.

Test of the restriction is as follows:

$$u = \frac{(SSR_{res} - SSR_{ols})/ J}{SSR_{ols} / n-k} \qquad (8)$$

where SSR_{res} and SSR_{ols} are residual sums of squares on the restricted and unrestricted (OLS) models respectively, J is the number of restrictions (deleted components) and n and k are as defined above. Because this is a test of the truth of $\underline{X}\underline{A}_2\underline{d}_2 = 0$, i.e., $\underline{A}_2 = 0$, the test statistic u is compared with a centrally distributed F, a classical F-test.

If one is willing to accept some bias for further reducing variances on $\underline{B}$, a mean square error (MSE) criterion can be used minimizing Σ(var + bias^2). This uses the same test statistic u but now comparison is with a non-central F distribution (Goodnight and Wallace 1972). Should further component deletion be desired, some eigenvalue size criterion might be employed, however, statistical properties of the resultant estimator, as in stepwise regression (Freund 1974) will be unknown.

Table 4. Results of the 1951 analyses under OLS, classical F, and MSE criteria of deer kill versus 35 land-use types defined in table 1.

Land-use type	$\underline{b}$ OLS	$\underline{b}$ Cl "F"	$\underline{b}$ MSE
AF	1.75 **	1.81 **	1.70 **
PS	-.269 **	-2.56 **	-2.43 **
UL	-.378 **	-3.60 **	-3.45 **
H2A	-0.37 ns	-0.08 ns	1.14 **
H2B	-3.96 **	-3.60 **	-3.59 **
H3A	-1.95 ns	-0.36 ns	-1.63 **
H4A	-3.80 **	-3.64 **	-3.05 *
S4A	9.31 *	9.27 *	9.12 *
HS1	-2.68 **	-2.18 **	-1.94 **
HS2A	-2.23 *	-3.22 **	-2.97 **
HS3A	-0.39 ns	0.91 ns	2.17 **
SH2A	1.12 ns	2.18 **	0.66 ns

* P < 0.05, ** P < 0.01, ns = nonsignificant

Results

The data sets for 1951 and 1971 were analyzed by principal component regression (tables 4 and 5). In the analysis of 1951 data, one component was deleted under the classical F criterion and one additional component deleted under the MSE criterion. For the 1971 data all components were retained under the classical F criterion, so the OLS model is the only unbiased estimator of $\underline{B}$ in terms of the specified model assumptions. Under the MSE criterion a single component was deleted.

Coefficients produced in the study are consistent from year to year and within year-period under OLS and both deletion criteria. Those effects found significant in both years, PS, UL, H2A, H3A, HS2A, and HS3A remained of the same sign and magnitude. The close agreement over the 20-year period and under deletion criteria within each year lead me to conclude that real relationships have been estimated.

Clearly, results show urban land as a significant negative effect in 1951 and 1971; this was to be expected. Other significant negative effects for both years were pasture, cutover hardwoods (H2B), older hardwoods (H3A), and a young mixed type (HS2A).

Table 5. Results of the 1971 analyses under OLS and MSE criteria of deer kill versus 35 land-use types defined in table 1.

Land-use type	$\underline{b}$ OLS	$\underline{b}$ MSE
PS	-2.04 **	-1.85 *
UL	-4.63 **	-4.26 **
H1	2.01 *	2.41 **
H2A	1.72 *	2.27 **
H2B	-3.87 **	-4.04 **
H3A	-1.72 **	-1.83 **
S1	-4.09 *	-4.03 *
S2A	4.30 **	4.84 **
S3A	-2.64 *	-2.18 *
S3B	3.65 *	3.87 *
HS2A	-1.62 ns	-2.52 *
HS3A	-0.33 ns	0.88 **
SH3A	-2.69 ns	-2.49 **

* P < 0.05, ** P < 0.01, ns = nonsignificant

Significant positive effects in both periods were young, dense-canopy hardwoods (H2A) an older, dense-canopy mixed forest (HS3A). Softwood types were found significantly positive in both periods but specific types were also identified as negative effects in 1971.

Age, canopy-closure and softwood mixture of hardwood and hardwood-dominant mixed forest types appear to be governing factors in terms of individual type effects on deer harvests. With softwood composition, this accounts for much of forest type effect on harvests. Young, dense hardwood stands are used by deer as a major source of browse while some of the more open softwood stands may function both as winter cover and feeding areas.

Those types that were negative (H2B, H3A, H4A, HS1, HS2A) may provide neither enough food nor enough escape cover in limited areas to be acceptable to deer. The negative softwood types (S3A, SH3A) because of their year-round dense canopies, likely provide little food or cover at all. Old softwood types that were positive effects (S3B, S4A) may function primarily as wintering sites with enough light penetration through or underneath the canopy to support an understory.

It is possible, therefore, to explain why the land-use types investigated had their calculated effects on deer harvests, although this may be viewed as a simplistic causal assumption. There are, of course, many effects of different land-use and forest types, both direct and indirect, that enter into the investigated relationships. Nevertheless, quantification of these general relationships may eventually lead to better definition of specific hypotheses about impacts of land-uses and forest composition.

CONCLUSIONS

Principal component analysis has shown an underlying structure in land-use and forest type composition in Massachusetts that is consistent over a 20-year period. Relationships between these axes of variability and deer harvest are also consistent and are capable of reasonable biological interpretation.

Principal component regression has allowed elucidation of individual type effects on deer harvest by reducing variances on $\underline{B}$ under criteria that produce known statistical properties. This is in direct contrast to data-dredging techniques, such as stepwise regression, that eliminate independent variables through mechanical manipulation and that leave $\underline{B}$ estimators that are unreliable, at best, because their statistical properties are not known. PCA and PCR are therefore seen as valuable multivariate analytical tools for the ecological investigator.

ACKNOWLEDGMENTS

I wish to thank Dr. Wendell E. Dodge of the Massachusetts Cooperative Wildlife Research Unit (U.S. Fish and Wildlife Service, Massachusetts Division of Fisheries and Wildlife, University of Massachusetts, Amherst, and The Wildlife Management Institute), Dr. Bernard Morzuch of the Department of Food and Resource Economics, University of Massachusetts, and Chet McCord and James McDonough of the Massachusetts Division of Fisheries and Wildlife for their advice and encouragement in this project.

LITERATURE CITED

Austin, M.P. 1968. An ordination of a chalk grassland community. Journal of Ecology 56(3):739-757.

Fomby, T.B., and R.C. Hill. 1978. Deletion criteria for principal components regression analysis. American Journal of Agricultural Economics 60(3):524-527.

Freund, R.J. 1974. On the misuse of significance tests. American Journal of Agricultural Economics 56(1):192.

Goodnight, J., and T.D. Wallace. 1972. Operational techniques and tables for making weak MSE tests for restrictions in regressions. Econometrica 40(4):699-709.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Johnston, J. 1972. Econometric methods. Second edition. 437p. McGraw-Hill, New York, N.Y.

Labisky, R.F., J.A. Harper, and F. Greeley. 1964. Influence of land-use, calcium and weather on the distribution and abundance of pheasants in Illinois. 19p. Illinois Natural History Survey Biological Notes No. 51.

MacConnell, W.P. 1973. Massachusetts mapdown: land use and vegetative cover mapping classification manual for use with Massachusetts mapdown maps. 19 p. University of Massachusetts Cooperative Extension Service Publication 97.

MacConnell, W.P. 1975. Remote sensing 20 years of change in Massachusetts, 1951/52 - 1971/72. 79 p. Massachusetts Agricultural Experiment Station Bulletin 630.

McDonough, J.J., and J.J. Pottie. 1979. A successful antlerless deer hunting permit system in Massachusetts. Transactions Northeast Fish and Wildlife Conference 36:110-119.

Nichols, S. 1977. On the interpretation of principal components analysis in ecological contexts. Vegetatio 34(3):191-197.

Page, G. 1976. Quantitative evaluation of site potential for spruce and fir in Newfoundland. Forest Science 22(2):131-143.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58(4):810-819.

Steel, R.G.D., and J.H. Torrie. 1960. Principles and procedures of statistics. 481 p. McGraw-Hill, New York, N.Y.

AN APPLICATION OF FACTOR ANALYSIS IN AN AQUATIC HABITAT STUDY[1]

T.J. Harshbarger[2], and H. Bhattacharyya[3]

Abstract.--In five small, high-gradient trout streams in western North Carolina, 18 cover variables were related to standing crop biomass of wild brook trout (Salvelinus fontinalis), rainbow trout (Salmo gairdneri) and brown trout (Salmo trutta) in randomly selected stream sections. Factor analysis of the data set showed that only a small number of factors or variables was needed to explain relations between variables in the observed set. Key cover factors were area in debris; turbulent water; vegetation, both in and over stream; and overhanging banks. Resolutions obtained were used in stepwise regressions to explore relationships between standing crop of trout and age of fish. Regressions containing factors as independent variables explained less variation in fish standing crop than did regressions containing equal numbers of original habitat attributes as independent variables.

Key words: Aquatic habitat; cover; factor analysis; multivariate analysis; regression analysis; stream fish; trout.

INTRODUCTION

The presence of a self-sustaining population of fish usually indicates compatibility between the aquatic envrionment and the ecological requirements of the fish. Wild trout are excellent indicators of current environmental conditions and their population density in streams reflects their level of compatibility With a highly integrated chemical, physical and biological situation. However, simply realizing that the trout population reflects its envrionment is not particularly informative to the resource manager. Explicit relationships between species and their environments are needed to assess the actual and potential capabilities of the habitat.

Functional and correlative approaches have been used to study factors influencing the distribution and abundance of a species. The functional approach is used when factors are known to influence certain attributes of the species. Correlative procedures are best suited for exploratory studies, where the relationship between a species and its environment are unknown. This procedure provides little information about causality, but helps the researcher to make inference for rigorous testing. Many variables can potentially influence the distribution and density of wild trout in a stream. Often the choice of parameters to measure and analyze is difficult because environmental variables in lotic waters are typically correlated and confounded with one another (Reid 1961).

Attempts to correlate single and multiple variables to trout populations in streams have met with varying degrees of success. Boussu (1954), Saunders and Smith (1962), and Wickham (1967) investigated the relationship of a trout

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Aquatic Ecologist, Southeastern Forest Experiment Station, Asheville, NC 28806.

[3]Mathematical Statistician, Southeastern Forest Experiment Station, Research Triangle Park, NC 27709.

population to cover. In two studies (Schuck 1943, Hunt 1969), water depth strongly influenced trout densities in streams, and current velocity was similarly implicated in another study (Lewis 1969). Few investigators, notably Lewis (1969), Stewart (1970), Platts (1974), and Binns and Eiserman (1979), have studied the simultaneous effect of several environmental variables on stream trout populations using multiple regression and correlation analysis.

The present study was designed to look at the relationships between various habitat parameters and trout. Factor analysis was used to delineate and examine a group of environmental variables that seemed important to trout, those providing cover or shelter. Regression analysis was used to examine relationships between the environmental factors produced and trout.

STUDY SITES AND METHODS

We selected five small, high-gradient trout streams in western North Carolina. In each, we inventoried and measured 18 variables providing fish cover, in 20 randomly selected 30m stream sections. Each section was surveyed along line transects established every 3m across the stream, perpendicular to its center line. All physical structures providing shelter or concealment for trout were located and their cross-sectional areas parallel to the air-water interface recorded. Structures included rocks or ledges which afforded cover to fish, undercut banks, aquatic vegetation, and logs and brush in the stream. Other cover situations such as water with sufficient surface turbulance to prevent visibility of the stream bottom were also measured and recorded.

Brush and loosely compacted debris in the stream and streamside vegetation overhanging the water surface were estimated ocularly. Cover afforded by brush and debris was recorded as the surface occupied by solid material and expressed as a percentage of the water surface it covered. Cover provided by overhanging bank vegetation was expressed in two variables: as the percentage of stream covered by vegetation between the water surface and 1.0m above, and the percentage between 1.0m and 2.0m above the surface.

Standing crop biomass of wild trout was estimated by depletion analysis electrofishing. All fish were weighed, measured, and returned to the midpoint of the section under study.

Factor Analysis

Starting with the observed correlation matrix on the 18 cover variables, we used SAS FACTOR programs for principal axis, maximum likelihood, and iterated principal axis factoring (SAS 1979). Using the principal axis method and keeping only those factors corresponding to eigenvalues greater than one, six factors were retained. To make the factors more conceptually meaningful, several rotational methods were used, including varimax, quartimax, equimax and promax (oblique) procedures (SAS 1979).

All factor and rotation procedures used on the 18 cover variables produced essentially the same information about underlying structure. However, the iterated principal axis and oblique rotational procedures are intuitively appealing. These procedures address communalities directly and recognize that some factors may very likely be correlated. As such only the interated principal axis solution using an oblique rotation of the cover related factors will be presented in this paper.

Regression Analysis

Relationships between standing crop of trout and the factors obtained from oblique rotation of the iterated principal axis solution were explored using SAS STEPWISE regression procedures and the maximum R^2 improvement technique developed by Goodnight (SAS 1979). Relationships between the habitat variables themselves and standing crop of trout were similarly explored to determine if factors entered the stepwise model in about the same order as the variable they contained. Coefficients of determination were compared to assess performance of models containing factors and variables.

RESULTS

The six factors combined the 18 cover variables into groups which generally reflected meaningful patterns in relation to the stream environment. Factors were named after the variable or variables producing greatest correlation. Factor loadings for the variables are shown in table 1. Factors had zero or close to zero projections on most variables, very few immediate loadings, and two or three high associations.

Factor 1 is a measure of debris with high loadings for number of logs and the surface areas, parallel to the stream, in logs and in brush. Factor 2 is a measure of side stream cover and is strongly negatively correlated with area in and percent cover of vegetation trailing in the water surface. The third factor expresses the percentages of cover provided by overstream vegetation 0-1m and 1-2m above the stream. Factor 4 is highly correlated with area in turbulent water and to a lesser degree with the number of units of turbulent water. This factor is also correlated with total cover and can be considered a general cover factor. Rock area and number load heavily on factor 5, and the sixth factor is highly correlated with area in overstream vegetation in the 1- to 2-m zone.

Factor values for each of the 100 stream section observations were obtained from the scoring coefficient matrix. Trout, regardless of

Table 1. Variables associated with cover factors.

Variables	Factor loadings 1	2	3	4	5	6
Ledge area (1)	0.063	0.241	0.030	0.400	-0.288	0.106
Rock						
Number (2)	-0.075	-0.047	-0.009	-0.096	0.850	0.214
Area (3)	0.099	0.194	0.114	0.109	0.646	-0.191
Turbulent water						
Number (4)	-0.108	0.304	-0.075	0.491	0.007	0.156
Area (5)	-0.244	-0.054	-0.059	0.936	-0.115	-0.014
Logs						
Number (6)	0.913	-0.049	-0.007	-0.108	-0.146	-0.003
Area (7)	0.969	0.003	-0.060	-0.089	-0.019	0.020
Bank area (8)	0.080	-0.539	-0.039	-0.125	-0.119	-0.080
Other area (9)	0.045	0.002	0.345	-0.051	0.112	-0.090
Brush						
% Cover (10)	0.102	-0.150	0.282	0.062	0.033	-0.012
Area (11)	0.824	0.076	0.067	-0.006	0.084	0.076
Side-stream vegetation						
% Cover (12)	-0.116	-0.772	0.048	-0.087	0.082	0.198
Area (13)	0.001	-0.712	0.006	0.012	0.011	0.015
Over-stream vegetation (0-1m)						
% Cover (14)	-0.045	-0.042	0.896	0.034	-0.109	-0.201
Area (15)	0.174	0.217	0.111	0.231	-0.333	0.363
Over-stream vegetation (1-2m)						
% Cover (16)	-0.032	0.089	0.844	-0.110	0.039	0.293
Area (17)	0.048	-0.050	-0.062	0.046	0.094	0.999
Total cover area (18)	0.365	-0.105	0.062	0.775	0.225	-0.076

species, in each stream section were segregated into the four age classes represented in each of the five streams under study. The relationship between factor values and trout standing crop for each age group was examined using the stepwise maximum R^2 regression technique. Likewise, the relationship between the original 18 cover attributes and trout standing crop was similarly examined.

The coefficient of determination (R^2) corresponding to the six factors was lower than that corresponding to the original 18 variables for each age class of fish (table 2). Further, the best one-attribute, two-attribute, etc., models obtained by the stepwise procedure also produced higher R^2 values when the original variables were used than when the deduced factors were used. Six-factor models produced coefficients (R^2) which ranged from 0.09 for the standing crop of the young-of-the-year trout to 0.53 for trout in age group II. Regressions on the original variables produced coefficients between 0.31 and 0.71 for 18 variable models and 0.26 and 0.66 for models containing six variables.

For young-of-the-year trout a single variable, rock area, produced an R^2 value (0.10) equivalent to that obtained from the model containing all six factors. In age group I, two habitat variables, numbers of rocks and total cover, produced a coefficient (0.18) as large as the six-factor model. A coefficient similar to that produced by the six-factor model was obtained for trout in age group II with three variables (R^2=0.56); number of rocks, percent cover provided

Table 2. Order in which the habitat attributes shown in table 1 entered stepwise regressions and the coefficient of determination they produced for each age group of trout.

Age group	Step 1 Attributes (R^2)	Step 2 Attributes (R^2)	Step 3 Attributes (R^2)	Step 4 Attributes (R^2)	Step 5 Attributes (R^2)	Step 6 Attributes (R^2)
Group 0						
Factors	6 (0.03)	4,6 (0.05)	3,4,6 (0.07)	2,3,4,6 (0.08)	1,2,3,4,6 (0.09)	1,2,3,4,5,6 (0.09)
Variables	3 (0.10)	2,3 (0.16)	2,3,12 (0.23)	3,12,15,27 (0.29)	2,3,12 15,17 (0.31)	2,3,10,12, 15,17 (0.32)
Group I						
Factors	5 (0.08)	4,5 (0.12)	1,4,5 (0.14)	1,2,4,5 (0.15)	1,2,3,4,5 (0.15)	1,2,3,4,5,6 (0.15)
Variables	2 (0.14)	2,18 (0.18)	1,2,10 (0.21)	1,2,3,10 (0.23)	1,2,3,4,12 (0.25)	1,2,3,4,10,12 (0.26)
Group II						
Factors	6 (0.17)	5,6 (0.40)	4,5,6 (0.46)	2,4,5,6 (0.49)	1,2,4,5,6 (0.52)	1,2,3,4,5,6 (0.53)
Variables	2 (0.31	2,3 (0.44)	2,10,17 (0.55)	2,10,15,17 (0.61)	2,10,12,15,17 (0.64)	2,10,12,15, 17,18 (0.66)
Group III						
Factors	6 (0.16)	5,6 (0.27)	2,5,6 (0.31)	1,2,5,6 (0.35)	1,2,4,5,6 (0.36)	1,2,3,4,5,6 (0.37)
Variables	17 (0.23)	15,17 (0.29)	12,15,17 (0.35)	4,12,15,17 (0.38)	2,4,12,15,17 (0.40)	2,4,10,12, 15,17 (0.42)

by instream brush, and area in overstream cover between 1 and 2 m; and for trout in age group III with four variables (R^2=0.34); number of pockets of turbulent water, percent cover of side-stream vegetation, area in overstream cover to 1 m, and area in overstream cover to 2 m.

Factor 5 (rocks) and factor 6 (area in overstream cover 2 m and above) and their equivalent habitat attributes were, in most instances, the first parameters to enter stepwise regressions containing factors or variables. Factor 2 (overstream cover) and factor 4 (general cover) entered models intermediately, and factor 1 (debris) and 3 (percent overstream cover) entered models last. Area in and percent cover provided by side-stream vegetation consistently entered models intermediately. Other habitat attributes either entered models intermediately or late in the stepwise procedure.

DISCUSSION

Factor analysis was useful in defining the underlying structure of the cover portion of the habitat for trout. Although 18 variables were singled out and measured, cover consists essentially of a six-dimensional space characterized by six factors: debris, side-stream cover, percent of overstream vegetation, turbulent water, rock area, and area of overstream vegetation in the 1- to 2-m zone.

In examining the relationships between standing crop of fish and the habitat attributes by stepwise regression methods, a higher coefficient of determination (R^2) was obtained by using the 18 original variables than by using the six derived factors. Moreover, it was found that the best one-attribute, two-attribute, etc., models also resulted in higher R^2 values when original variables were used. This indicates that when the original variable measurements are available, there is no reason to form regression models based on derived factors. With the exception of 2-year-old trout, no R^2 value between trout and the set of variables exceeded 0.5, indicating that there is a substantial portion of variability in fish biomass that is not explained by the measured variables and hence also not explained by the derived factors. Part of this variability may be accounted for by water or flow related variables. The results of a combined analysis of water and cover variables will be reported at a later date.

LITERATURE CITED

Boussu, M.F. 1954. Relationship between trout populations and cover on a small stream. Journal of Wildlife Management 18:229-239.

Binns, N.A., and F.M. Eiserman. 1979. Quantification of fluvial trout habitat in Wyoming. Transactions of the American Fisheries Society 108:215-228.

Hunt, R.L. 1969. Effects of habitat alteration on production, standing crops, and yield of brook trout in Lawrence Creek, Wisconsin. p. 281-312. In T.G. Northcote, editor. Symposium on salmon and trout in streams. H.R. MacMillan Lectures in Fisheries, Vancouver, Canada.

Lewis, S.L. 1969. Physical factors influencing fish populations in pools of a trout stream. Transactions of the American Fisheries Society 98:14-19.

Platts, W.S. 1974. Geomorphic and aquatic conditions influencing salmonids and stream classification--with application to ecosystem classification. USDA Forest Service, Intermountain Forest and Range Experiment Station, Boise, Idaho.

Reid, G.K. 1961. The ecology of inland waters and estuaries. 375 p. Reinhold, New York, N.Y.

SAS. 1979. Users guide. 494 p. SAS Institute, Inc., Raleigh, N.C.

Saunders, J.W., and M.W. Smith. 1962. Physical alteration of stream habitat to improve brook trout production. Transactions of the American Fisheries Society 91:185-188.

Schuck, H.A. 1943. Survival, population density, growth, and movement of the wild brown trout in Crystal Creek. Transactions of the American Fisheries Society 73:209-230.

Stewart, P.A. 1970. Physical factors influencing trout density in a small stream. Ph.D. Dissertation, Colorado State University, Fort Collins, Colo.

Wickham, M.G. 1967. Physical microhabitat of trout. M.S. Thesis, Colorado State University, Fort Collins, Colo.

DISCUSSION

TERRY LARSON: Most papers given here which involved multiple regression analysis used stepwise techniques. Why did you not use an all possible subsets technique like BMDP-9R? This program is not costly to run and will pick up subsets that stepwise techniques miss.

HELEN BHATTACHARYYA: The SAS STEPWISE procedure with MAXR option was used. You are quite right that all possible regression (SAS RSQUARE procedure) may pick out combinations not covered by stepwise. However, STEPWISE/MXR is almost as good (see SAS writeup) and has the advantage of printing all other regression statistics, slope, intercept, mean squares, etc., besides just the R^2 value.

New Approaches to Analysis and Interpretation

BIRD COMMUNITY USE OF RIPARIAN HABITATS: THE IMPORTANCE OF TEMPORAL SCALE IN INTERPRETING DISCRIMINANT ANALYSIS[1]

Jake Rice[2], Robert D. Ohmart[3], and Bertin Anderson[4]

Abstract.--Discriminant functions analyses were used to differentiate habitats used from habitats not used by every bird species in the lower Colorado River riparian areas, during the breeding season. Most of these DFA's produced statistically significant results, but the mean percent of transects correctly classified into the species-present or species-absent groups, across all species, was less than 75%. Patterns of errors of classification were examined and found to be random with regard to vegetation community, foliage structure or avian species.

When we considered a second year's census data, for 18 of 21 species a significant number of errors in predicted occurrences were "corrected" with the new avian distributions. The habitat DFA's were then repeated between the group of transects used both years and those not used both years. Across all species and all seasons, transects could be correctly classified with an accuracy of 89%. We also found biologically meaningful patterns of seasonal variation 1) in habitat selectivity of the avian community, 2) in differences among vegetational communities and 3) of transects with irregular occurrences of the avian species.

Key words: Birds; Colorado River; community ecology; discriminant function analysis; habitat use; riparian; species turnover.

INTRODUCTION

During the past decade there has been a major upswing in the use of multivariate statistics in the study of ecology. In avian studies the uses have been largely to quantitatively describe community structure through various types of ordinations (James 1971, Whitmore 1975, Conner and Adkisson 1977), or to document niche partitioning among groups of species (e.g., Cody 1968, Hespenheide 1971, Whitmore 1977). This predisposition toward each of these goals can be understood, given current areas of emphasis of theory in community organization and evolutionary ecology in general. The relationships among bird

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop. April 23-25, 1980, Burlington, Vt.

[2]Research Professor, Department of Zoology, Arizona State University, Tempe, AZ 85281 and Assistant Professor, Biology Department, Memorial University of Newfoundland, Canada, A1B 2X8.

[3]Associate Professor, Department of Zoology and the Center for Environmental Studies, Arizona State University, Tempe, AZ 85281.

[4]Research Associate, Department of Zoology and the Center for Environmental Studies, Arizona State University, Tempe, AZ 85281.

species, and between bird species and characteristics of their habitats, have proven reconcilable with theory, and have often even led to enlightening extensions of our understanding of ecological processes.

On the other hand, a large number of these studies can be, and often have been, criticized (if rarely in print, frequently in discussions) for taking a very casual attitude toward several aspects of the original data. There are a variety of potential sources of variation that can be reflected in ecological data, and valid, useful, management studies require research designs that take adequate account of these kinds of variability. Specifically, relatively little attention has been given to verifying that the basic data sets on which the multivariate techniques are used: 1) sample the true and complete range of habitats acceptable to species included in an area; 2) cover the within-species variability in habitat use across communities; and 3) cover the season-to-season and year-to-year variation in bird-vegetation relationships, although a few papers considering some of these points can be found (e.g., Smith 1977, Rotenberry 1978, Rotenberry et al. 1979).

A comprehensive understanding of community ecology will require detailed investigation of each of those sources of variation. Additionally, when investigations move from the theoretical realm to the practical, all of these considerations become even more important. If habitat management plans are actually going to be developed and implemented based on the results of sophisticated quantitative studies, it is essential that the findings truly reflect the species-habitat relationships and are not just statistically significant or consistent with one of many diverse theoretical expectations.

As specific examples, the questions of which spatial scale and which temporal scale to sample become of paramount importance. Furthermore, the errors that occur, for example, in the classification step of a discriminant function analysis, are no longer simply inconveniences or embarrassments, but they become real problems affecting the potential success of any habitat management plan. This paper presents discriminant functions investigations into the habitat use patterns of an entire avifauna; looking particularly at what the errors of classification truly represent, at least in our system, and what temporal scale is suitable for bird habitat studies.

Riparian habitats in the desert Southwest are rich ecological oases for many species of birds and mammals. These areas are also subjected to intense competing land and water use demands for agricultural lands, municipal and industrial purposes, river channel and flood control, and recreational use, in addition to their value for wildlife. For several years the Colorado River Project has been quantifying wildlife densities and use of all riparian habitats along the lower Colorado River. We are currently using these data to develop a predictive model of bird-vegetation (and soon mammal-vegetation) relationships within the system. This model will be used by state, federal and private concerns in actual land use decision making and will be a major tool in assessing and planning habitat mitigation in these riparian areas.

Because of the intended use of our findings we had to know not simply how species X and Y differed in habitat preferences in an area, or even what major gradients we could uncover in community structure. Rather, we had to know, out of the complete range of riparian habitats available, which habitat factors determined or allowed the occurrence of each avian species. Our answers had to be valid over the entire year, because management decisions directed toward a single season will nonetheless have year-round ramifications. Correspondingly, our results had to be valid for several years, not for just whatever special conditions reigned for any single year.

METHODS

It required 72 transects of 1600 m or 900 m to census every community present in each stage of development and in the proportion in which each habitat type occurred along the lower Colorado River. Each transect was then censused three times each month using the familiar strip method (Emlen 1971, 1977). On the basis of climatic and demographic patterns, we divided the year into five seasons: Spring (March-April), Summer (May-June-July), Late Summer (August-September), Fall (October-November) and Winter (December-January-February). Censuses within each season were averaged when we calculated bird species occurrences and densities.

Every tree by species within 16 m of each side of each transect was counted. Also foliage density measures at several heights were taken at 50 m intervals along each transect. The vegetation density measures were combined to produce, for each transect, a 12 variable array describing the foliage density at each stratum, the relative species composition (transformed with the ARCSIN square root transformation), and the foliage height diversity of that transect. These variables were all habitat attributes which we believed were both potentially adequate to characterize habitat use by members of the avian community and were realistically manageable from both the standpoints of data analysis and habitat management.

We used discriminant functions analyses to quantify habitat use attributes of each species. For each species, each season, we divided the pool of 72 transects into two groups: those where the species in question was recorded, "Present", and those where the species was not recorded, "Absent". The Present and the Absent groups of transects were then differentiated on the array of

habitat measures, to quantify both how different the used and the unused areas were vegetationally, and what attributes of the vegetation characterized the areas used by each species in the community. Actually, for many species (those with widespread distributions and widely differing abundances on different transects), we repeated the analyses using three or even four groups, based on increasing densities. The accuracy of those analyses were comparable to those of the two group discriminations. The multigroup analyses only introduce further complications because of the additional possible axes of discrimination, and they will not be discussed.

RESULTS AND DISCUSSION

Single-year Discriminations

A synopsis of results of discriminant function analyses for the 39 species which were present during the summer is presented in figure 1. The majority of individual species discriminations showed that habitats were significantly different between used and unused areas. However, quite a number were not different statistically; a matter for possible concern. Some of these cases may represent strictly statistical problems, of the sort discussed by Williams (1981), whereas others might represent species truly showing no vegetation differences between used and unused sites.

Rather than go into all possible investigations of sources of statistical artifacts or errors, we had previously decided that results of the classification step of the discriminant function analyses would be most appropriate for our model, and to us it represented the most important measure of the success of the analyses. We knew, for example, that necessary assumptions of homogeneity of variances and covariances between the Present and Absent groups of transects would often be violated by species with either very widespread or restricted occurrences in the riparian vegetation. Such violations would affect the statistical significance of the analyses. However, regardless of the statistical significance of any given discrimination, if we were able to correctly identify areas that were commonly used by a particular species on the basis of vegetation attributes, we felt we would have a useful management tool.

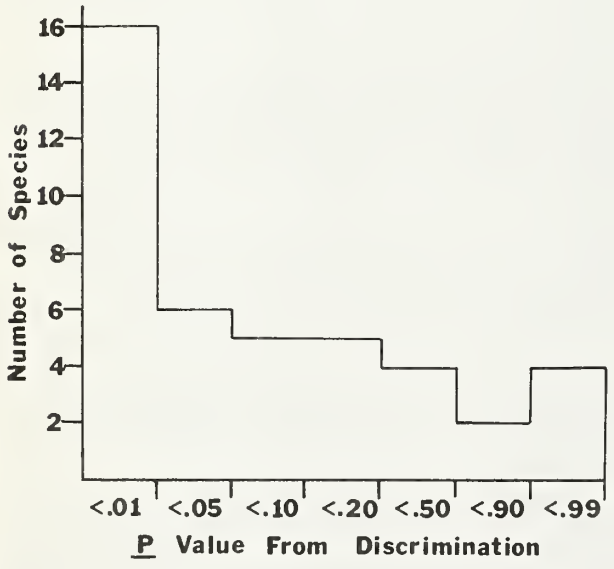


Figure 1. The number of species for which the discriminant analyses of used versus unused habitats reached specific levels of significance. Distribution data were from a single summer.

Table 1. Types of transect classifications possible and their biological significance.

		Status predicted from classification step of the discriminant analyses.	
		PRESENT	ABSENT
Actual status of species on transect	PRESENT	A	B
	ABSENT	C	D

A = Correct prediction of the species presence on the transect

D = Correct prediction of the species absence on the transect

B = Suitable habitat identified as unsuitable for the species

C = Unsuitable habitat identified as suitable for the species

Statistically, the classification step of a discriminant function analysis could produce two different types of errors, illustrated in table 1. From a management standpoint, the two possible types of classification errors are quite different in their importance, and use determines which errors are most serious. To include some unused areas in the used group (C errors, table 1) would be a conservative error for habitat preservation activities. It would lead to preservation of some unsuitable habitats as well as all suitable ones. The other type of error (B errors, table 1) would be more serious; one would reject areas which were, in fact, good habitats for the species in question. Conversely, if one were engaged in management activities to create or modify habitats, the seriousness of the errors would be reversed. To make type B errors would be to

develop only some of the range of habitats suitable for a species, whereas to make type C errors would be to spend time and resources developing habitats which would turn out to be unsuitable for the species.

Analysis of Errors

Figure 2 shows the distribution of the percent of correct classifications of transects for the single year discriminant analyses. Although we found few gross failures in the classification steps, there were also correspondingly few species for which we had great success at predicting transect suitability. Rather than continue to use this equivocal tool with data from other seasons, we decided to look in detail at what sorts of errors were occurring, in hope of isolating specific problems of the approach. Possible sources of errors in this study (and correspondingly, other similar studies) included: 1) uneven variability in either the suitability of habitats or the distribution of the species on a scale small enough to affect our findings; 2) inadequate habitat measures, i.e., we had not measured plant community traits important to avian habitat selection; and 3) low habitat selectivity by the bird species, i.e., the used and unused areas truly did not differ. To be tractable, we chose 21 of 39 avian species for detailed investigation, arbitrarily selecting the first 21 species from an alphabetical listing of all the species present.

We looked first at the habitat variability problem; that is, were our discriminations poor because some specific habitats were particularly good or bad at supporting avian species? This is an aspect of the question posed by several ecologists, such Colwell and Futuyma (1971) and

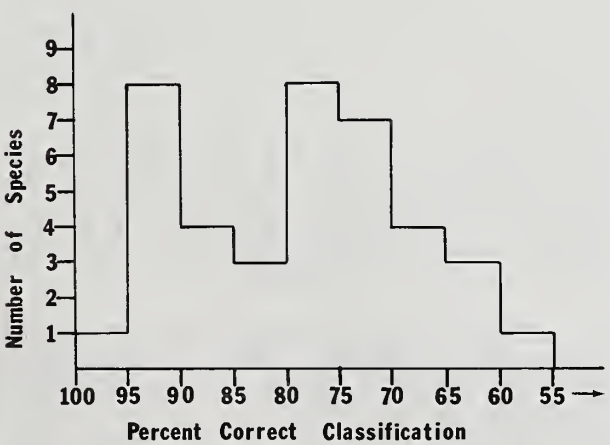


Figure 2. The number of species for which the classification functions of the discriminant analyses were able to correctly identify transects as used or not used at various accuracy rates.

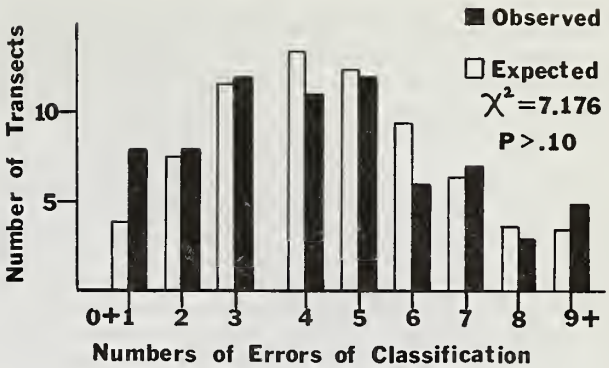


Figure 3. Fit of number of errors of classification of species suitabilities per transect to a Poisson distribution.

Willson (1974); that is, are all equal mensural discontinuities along a resource continuum of equal biological importance?

We investigated this problem in a simple way. Across the 72 transects there was a mean of 4.64 species misclassified per transect. If these errors were random, we would expect the number of transects with no species misclassified, with one species misclassified, with two species misclassified, and so on, to be distributed as a Poisson (random pattern) with a mean of 4.64. If specific types of habitats were either better or worse than average in terms of their ability to support species, relative to other habitats nearby on the vegetation continua represented by the discriminant functions, the distribution of errors per transect would deviate from the expected values. There was an excellent fit to the predicted distribution (fig. 3). From this we concluded that our error rate was not tied, at least primarily, to differential habitat attractiveness, beyond those differences captured in our vegetation measurements on each transect.

Initially we had intended to use data from subsequent years to test the effectiveness of the model. However, looking at the distributional data from the next year might shed light on the degree of consistency of habitat selection of the species in the avian community. From the classification step, we knew for each species: 1) the number of transects where the species was absent, yet were classified by the analysis as suitable habitat, and 2) the number of transects where the species was present, yet were classified by the analysis as unsuitable habitat. From the occurrence data for the next summer across the same 72 transects, we also determined: 3) the number of transects missing each species in the first year but supporting it in the second; and 4) the number of transects supporting each species in the first year but missing it in the second. All four of these counts can be converted into probabilities simply by dividing by 72 (the number of transects). If errors of classification were true errors with no biological relationship to

species occurrences, the product of 1 and 3 and of 2 and 4 would give the expected number of new species occurrences that were "corrections" of previous type C errors and the number of new species absences that were "corrections" of previous type B errors.

We found a surprisingly high rate of species turnover. New appearances occurred on 18% of all possible species-transect combinations, and new absences occurred in 16% of the species-transect combinations. On a species-by-species basis, the predicted number of independent "corrections" due to these species turnovers was usually too low for statistical comparison (commonly one to three expected "correct" new transect appearances or absences per species). However for 18 of the 21 species, the actual number of correct new appearances was greater than the predicted number, and for 16 of 21 species, the observed number of correct new absences was also greater than the predicted number. Using a binomial test, we determined that both of these divisions were statistically significant. Therefore, a significant number of what appeared to be errors of classification based on one year's distributional data were actually valid predictions of future distributions of the species.

Adding data from a second year resulted in a trade of one problem for another. In addition to some new occurrences being in areas previously identified as suitable, other new occurrences of each species were in transects previously classified correctly as unsuitable for that species. Absences in the second year were also noted in areas which previously had been classified correctly as suitable with data from only one year. Although these new errors were usually less frequent than expected by chance, they still outnumbered the "corrections" for both kinds of changes in the status of bird species between the two years (table 2). Clearly, the major problem affecting our ability to define quantitatively and precisely the range of acceptable habitats for each species was the high rate of species turnover from year to year on the same transects.

Table 2. Rates of species turnover per transect, and changes which were "corrections" or new errors of status relative to discriminant analysis classification.

Overall mean probability of status change from:			
Year 1	to	Year 2	
Absent		Present	0.183
Present		Absent	0.164
Number of observed cases > number of predicted cases from independence of species turnover rates and errors of classification:			
Correct new presences:		18 of 21 species Binomial "P" = 0.0012	
Correct new absences:		16 of 21 species Binomial "P" = 0.018	
Number of new correct presences = 87			
Number of new errors of presence = 108			
Number of correct new absences = 49			
Number of new errors of absence = 87			

Table 3. The mean percent of transects which were correctly classified as supporting or not supporting each species by season:

Season	Group		Total
	Present	Absent	
Summer	89.7	84.6	86.5
Late Summer	94.8	89.1	91.3
Fall	91.1	87.1	88.5
Winter	87.7	86.7	87.1
Spring	89.3	91.2	90.0
Total	90.8	88.0	89.0

Two-year Discriminations

In light of the high rate of species turnover on the transects, for each species each season we regrouped the transects on a new criterion; consistency of species occurrence. Three groups were formed: 1) transects where species X was recorded both years, 2) transects where species X was absent both years, and 3) transects where species X was present in one of the years but absent in the other. Discriminant function analyses were then conducted between groups 1 and 2 for each species each season.

Results of these analyses showed a marked improvement in ability to predict correctly areas which would or would not be used by each species. Regardless of season, the mean percent of transects classified correctly was always high, and the important groups of transects used consistently were even more frequently identified correctly (table 3). This high rate of correct transect classification also eliminates the two other possible sources of error in the one-year discriminations. When investigated at an appropriate temporal scale, the bird species

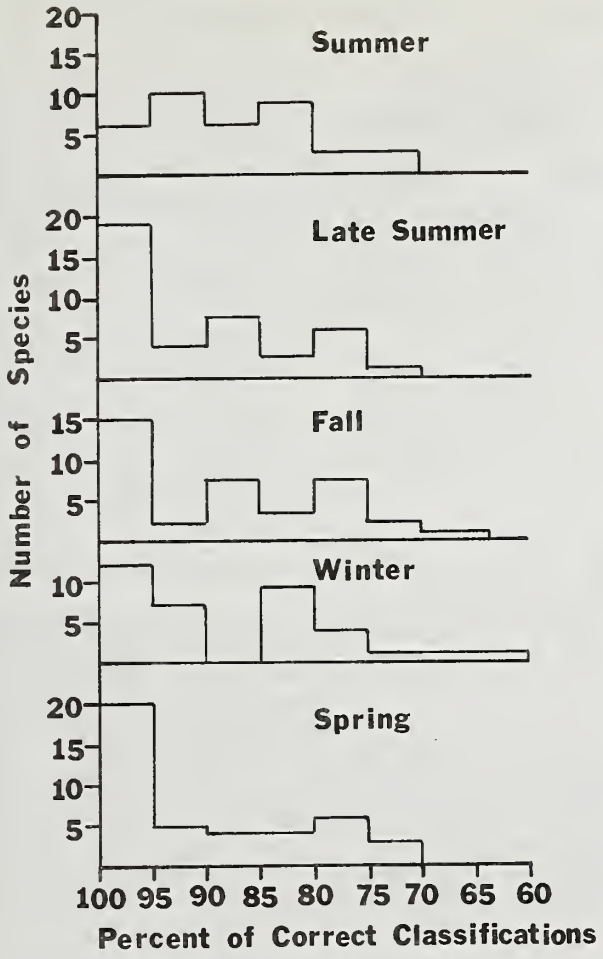


Figure 4. The number of species each season for which the classification procedures were able to correctly identify transects as used or not used with various accuracy rates. Distributional data were from two years.

generally did show habitat selectivity, and our habitat measures were adequate to detect this selectivity. The spread of the percent of species distribution predicted correctly by season (fig. 4) supports the impressions from table 3 and additionally underscores several points of general ecological significance.

One point apparent from figure 4 was the difference in the distribution of transect identifications by species between the summer and other seasons. The summer season had both the lowest mean transect identifiability and a distribution markedly more unimodal than the distribution of the transect identifications by species for the other seasons. This is, in itself, a noteworthy insight into the system we are studying.

The growing bimodality of the distribution (fig. 4) through late summer, fall, and winter is important for selecting a scale in which to undertake habitat management investigations. In summer nearly all species showed a fairly high degree of habitat selectivity, but almost none occupied areas perfectly or nearly perfectly discriminable from unoccupied areas. At other seasons the number of species with perfectly discriminable habitat-use patterns made up a substantial fraction of the community, whereas the habitat discriminability for other avian species in the community is markedly lower. The structure of the community changed from one of essentially all moderate habitat specialists in summer to a mixture of some extreme habitat specialists and other species showing quite weak habitat specificity through the fall, winter, and spring.

Considerable variability was still observed in the degree of habitat selectivity reflected by individual species. For resident species there was more variability between seasons in the discriminability of habitats used by each species (as reflected in the mean range of percent correct classifications) than between the rest of the community and a randomly selected species each season (table 4). Furthermore, the criteria of habitat selection shown by resident species, as reflected by variables with high loadings on the discriminant functions each season, changed as much as did the degree of habitat selectivity. Only 4 of 20 year-round residents had any one variable significantly different between used and unused areas for all five seasons (table 5). Information of this type is obviously of great value to persons involved in applied ecological work, and there are even a few substantial theoretical implications of these findings.

Remaining Errors

What of the errors of classification which remain: that 10% of the classifications which are still in error; and what of the distribution of the irregular transects on the discriminant axes of each bird species? Space does not allow an in-depth examination of both of these points, but we can consider the biologically most important one: areas used consistently by a species but nonetheless classified unsuitable for occupancy. In a management context, making such errors would involve the loss of habitat of high quality to the species of concern.

We again fitted the observed number of transects with no unpredicted species present, with one unpredicted species present, with two, three, and so on to the expected Poisson distribution for each season (fig. 5). In every season except winter the observed distribution deviated significantly from the predicted one, and in winter the fit was marginal. The preponderance of transects with no errors implied that the habitat suitability predictors, that is, the discriminant functions, were very accurate most of the time. Even more noteworthy, when we looked at those transects with three or four unpredicted

Table 4. Measures of the amount of variability of transects correctly classified for resident species.

Within species:	Between species:				
Mean maximum range among seasons	Mean range around a randomly selected species each season				
	Summer	Late Summer	Fall	Winter	Spring
$\bar{X}$ = 18.16%	11.58%	13.75%	8.89%	8.57%	8.81%
s.d.= 7.25	7.22	5.62	4.30	6.39	3.89

Table 5. Criteria of habitat selection as reflected by variables significant in stepwise discriminant analyses of present vs. absent transects for resident species all seasons.

Species	Summer	Late Summer	Fall	Winter	Spring
Verdin (Auriparus flaviceps)	None	None	None	None	None
Cactus wren (Campylorhynchus brunneicapillus)	HM+,SC,HMt[1]	FHD,SC	SM+,SC	HMt	SC
House finch (Carpodacus mexicanus)	W,HM+HMt	C,HMt,FHD, SC	None	C,SM+,SC	None
Gila woodpecker (Melanerpes uropygialis)	FHD,W,C	FHD,HM+,W	FHD,O,HM+, HMt	FHD,HM+, HMt, O	FHD,HM+,C, HMt,SC,W, V5,TV
Common flicker (Colaptes auratus)	W,FHD,C SM+	FHD,HM+,O	V5,SC	W,HM+	TV
Gambel's quail (Lophortyx gambelii)	SC,HM+, W, V15,C,HMt	FHD	SC,FHD	V15	V15,V5, FHD,C, SM+,SC
Ladder-backed woodpecker (Picoides scalaris)	FHD,V6,V15, HM+	None	None	None	V5
Roadrunner (Geococcyx californianus)	V5	FHD	None	None	HMt
Loggerhead shrike (Lanius ludovicianus)	SC,HM+,HMt, SMtV15	C,FHD	V15,V6,V5, HMt,O	FHD,W	O
Song sparrow (Melospiza melodia)	C,FHD,W, V15	FHD,HMt,W, C	FHD,C,SC,W, HMt,SMt,O	FHD,C,W, HMt	FHD
Mockingbird (Mimus polyglottos)	HM+,HMt,SC	HMt	HMt	HMt	HMt
Ash-throated flycatcher (Myiarchus cinerascens)	FHD,SC,HM+, O,HMt	None	HMt,FHD,V15, O,V5,V6,SM+	None	V6,TV,C
Phainopepla (Phainopepla nitens)	HMt,HM+,SC	HM+	HMt,SM+,V15	HMt,SM+	HMt

Abert's towhee (Pipilo aberti)	FHD	None	None	None	None
Black-tailed gnatcatcher (Polioptila melanura)	SC,HM+,C,W, HMt,V15,V6, TV	FHD,SC,W	FHD,W	FHD	TV,V6,C
Rough-winged swallow (Stelgidopteryx ruficollis)	W,C	None	None	None	None
Crissal thrasher (Toxostoma dorsale)	SC,HM+,HMt	FHD	FHD,W,V15, V5	FHD	V15,SC,C FHD,V6
Western kingbird (Tyrannus verticalis)	W	None	O	W	C,SM+,SMt
Mourning dove (Zenaida macroura)	FHD,HMt,HM+	None	None	None	V6,TV,V5, V15
White-crowned sparrow (Zonotrichia leucophrys)	HM+	None	HMt	FHD	V15,TV,HMt, W,HM+,SC

[1]Variable symbols:

V6 = Foliage volume 0.1 m - 0.6 m (0.5 ft - 2 ft)
V5 = Foliage volume 1.6 m - 3.1 m (5 ft - 10 ft)
V15 = Foliage volume 4.6 m and greater (15 ft)
TV = Total foliage volume
FHD = Foliage height diversity
HMt = Total proportion of honey mesquite (Prosopis glandulosa)
HM+ = Total proportion of honey mesquite with mistletoe (Phoradendron californicum)
SMt = Total proportion of screwbean mesquite (Prosopis pubescens)
SM+ = Total proportion of screwbean mesquite with mistletoe
SC = Total proportion of salt cedar (Tamarix chinensis)
W = Total proportion of willow (Salix gooddingii)
C = Total proportion of cottonwood (Populus fremontii)
O = Total proportion of mixed other species

species present, we found that certain types of communities displayed greater than chance frequency (table 6). Specifically, honey mesquite and/or screwbean mesquite were often classed as not suitable for species which did occur; i.e., their value to wildife is underestimated, often by as many as four or more species. This knowledge can be incorporated readily into habitat management plans.

CONCLUSION: TEMPORAL SCALE

To return to one of the major questions we initially posed, the discriminant analyses from one year's occurrence data provided a frequently significant but nonetheless weak ability to predict habitat use by each species in the riparian summer community. Merely incorporating a second year's data on distribution improved the accuracy of the predictive system markedly. For ecologists interested in the habitat attributes of species or communities, we would strongly recommend that however the initial habitat suitability judgments are formed, be they from line transects, singing male perches, from spot mapping, or whatever, a single year's data are not adequate for the study. The habitat-use patterns of avian species simply show too much year-to-year variability. Furthermore, the marked changes in community organization overall, and in habitat selectivity and selection criteria of individual species between seasons, indicate that a temporal scale smaller than an entire year is also inappropriate for a complete study of wildlife-habitat relationships.

In terms of extending the use of multivariate analyses, a stated purpose of this conference, we emphasize that many of the points made here did not come from a consideration of statistical significance of the analyses, nor from a consideration of the relative contributions of various habitat measures to the discriminant functions (the two things commonly looked at first). Rather they came from a careful consideration of errors made in the classification steps and attention to the adequacy of data to truly represent the system under study. As with many other lines of scientific inquiry, it was through attention to failure of the initial approach that subsequent insights arose.

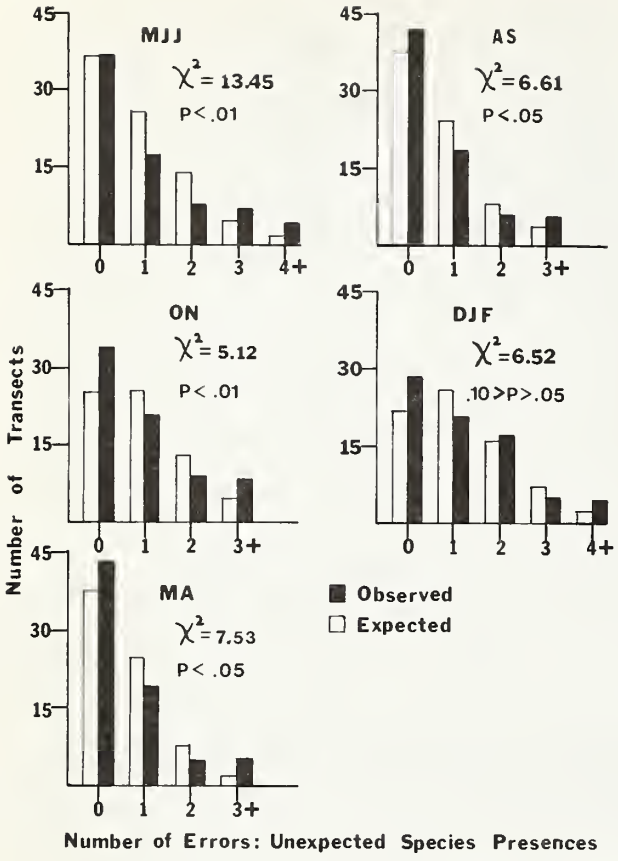


Figure 5. Fit of numbers of errors of classification of species occurrences per transect over two years to a Poisson distribution by season.

Table 6. Habitats where suitability for avian species is significantly often erroneously predicted by season.

Season	Suitability significantly often:	
	Underestimated	Overestimated
Summer	Screwbean mesquite	None
Late summer	Honey mesquite and screwbean mesquite	Salt cedar
Fall	Honey mesquite and structure type IV	None
Winter	None	None
Spring	Honey mesquite and screwbean mesquite (P = 0.063)	None

ACKNOWLEDGMENTS

We thank J. Anderson for criticisms of the manuscript. Cindy D. Zisner typed the final draft and prepared some figures. We are grateful to the many field biologists who helped collect the data. Kurt Webb carried out the computer analyses. This work was funded through Water and Power Resources Service Grant #7-07-30-V0009.

LITERATURE CITED

Cody, M.L. 1968. On the methods of resource division in grassland bird communities. American Naturalist 102:107-147.

Colwell, R.K., and D.J. Futuyma. 1971. On the measurement of niche breadth and overlap. Ecology 52:567-576.

Conner, R.N., and C.S. Adkisson. 1977. Principal component analysis of woodpecker nesting habitat. Wilson Bulletin 89:122-129.

Emlen, J.T. 1971. Population densities of birds derived from transect counts. Auk 88:323-342.

Emlen, J.T. 1977. Estimating breeding season bird densities from transect counts. Auk 94:455-468.

Hespenheide, H.A. 1971. Flycatcher habitat selection in the eastern deciduous forest. Auk 88:61-74.

James, F.C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin 83:215-236.

Rice, J.C. 1978. Ecological relationships of two interspecifically territorial vireos. Ecology 59:526-538.

Rotenberry, J.T. 1978. Components of avian diversity along a multifactorial climatic gradient. Ecology 59:693-699.

Rotenberry, J.T., R.E. Fitzner, and W.H. Rickard. 1979. Seasonal variation in avian community structure: Differences in mechanisms regulating diversity. Auk 96:499-505.

Smith, K.G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology 58:810-819.

Whitmore, R.C. 1975. Habitat ordination of passerine birds of the Virgin River valley, southwestern Utah. Wilson Bulletin 87:65-74.

Whitmore, R.C. 1977. Habitat partitioning in a community of passerine birds. Wilson Bulletin 80:253-265.

Williams, B.K. 1981. Discriminant analysis in wildlife research: theory and applications. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [April 23-25, 1980, Burlington, Vt.]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Willson, M.F. 1974. Avian community organization and habitat structure. Ecology 55:1017-1029.

DISCUSSION

B.K. WILLIAMS: I want to compliment you on this focus of discriminant analysis on its classification capabilities. I would suggest that we should generally shift our perspective on this methodology more toward classification and away from group mean separation.

JAKE RICE: Thanks, and I agree with your feelings about a change in emphasis.

MARTIN RAPHAEL: Did you use equal or prior probabilities in your analysis of classification success?

JAKE RICE: Priors.

BOB CLARK: Once you have determined that a species is present, can you use DFA to classify densities (low, moderate, high) or is the system too variable? How does your food availability data tie into presence/absence on "predicted" suitable habitat?

JAKE RICE: We have had limited success with density classification; about comparable to that of the initial one-year discriminations. As you suggest, year-to-year variability in abundance is the major problem. We are including abundance predictions in the model being constructed with these DFA results, but as a step after predicting species composition for a locality. We are using a regression approach to abundance predictions, and not surprisingly, the confidence intervals are large, because of the great year-to-year variability in abundance at the same sites.

As for food availability, we have the necessary data but the analyses are not yet far enough to provide useful answers to the problem of bird distributions. Not surprisingly, insect and seed abundances are at least as variable, seasonally and yearly, as are bird distributions. Insects and seeds are also spatially highly variable, and the variation is asynchronous between sites. Tying together two such variable systems is going to be a long, slow process.

MARK BOYCE: Since your characterization of habitat is based solely on vegetation characteristics, I am concerned that other components, especially insect abundance, may vary temporally, thus possibly invalidating your remarks regarding generalists vs. specialists.

JAKE RICE: My comments on generalists and specialists were meant to apply solely in terms of habitat selection attributes. Empirically the figure shows that in late summer, fall and winter some species in the community have their used habitats clearly differentiated from areas not used, whereas other species show little habitat differentiation. It was the species showing little differentiation that I was calling generalists. Much current ecological theory would predict that these habitat generalists would be specialists on some other criterion. EVERY species could be (and probably is) a specialist on some ecological attribute, but such an approach to ecology (i.e., studying every species until one found SOMETHING on which it specialized) would provide only a limited insight into community organization, overlooking as it would all ecological attributes where a species was more generalized as being uninteresting or "invalidated" by the finding that it was a specialist on something.

JAMES DUNN: Will you clarify your statement that stepwise variable selection often resulted in entirely different variables from one season to another. Is this mainly because your habitat variables are sensitive to seasonal and/or yearly change? Or does habitat preference actually change seasonally?

JAKE RICE: To a small extent, values of the habitat measures do change seasonally. That is the case only for foliage volume measures, of course, and not tree species composition measures. The changes in significant variables reflect, in very large part, changing distributions of the species by transect, and inferentially, changing criteria of habitat selection.

JAMES DUNN: Your attack on the problem suggests that you believe that site suitability for a species is not a simple yes/no question, but rather has a range of probabilities. If so, then why not use a classification method which works by assigning a probability for each site, e.g., the multivariate model as proposed by S.H. Walker and D.B. Duncan (1967. Estimation of the probability of an event as a function of several independent variables. Biometrika 54:167-179).

JAKE RICE: We have been moving in precisely that direction with our model; predicting which species will be found in a specified area with certainty, with high probability, often, etc. The classes are pretty rough, but adequate for our users. Thank you for the reference.

PAUL GEISSLER: You have suggested sampling the same transect in several years. Resampling transects provided very valuable information for your study. However, for the different objective of determining bird habitat relationships, I think it would be advantageous to take a new sample of routes each year to provide protection against the effect of some unmeasured and possibly unmeasurable habitat effect being confounded with the effects of measured habitat variables. To put it another way, the measurements on the same route have correlated residuals.

JAKE RICE: I do not really see how "determining bird habitat relationships" is a "different objective" from what we are attempting. The major

point here is that, from our data (and we think our findings are pretty general; surprisingly few data are available on the consistency of species occurrences and densities over several years at same sites) a single year's censuses are not reliable indicators of bird distributions, and therefore also are not reliable indicators of a species "habitat preferences" (or "optimal situations," if you prefer). You need multiple year's data to even get a good idea of a species' distribution pattern. Statistical considerations like correlated residuals are important, but come secondary to getting a reliable measure of the phenomenon one is trying to explain. Our point is that one year's data will not provide a reliable measure to predict, to discriminate or otherwise to statistically manipulate.

E. JAMES HARNER: Just a comment. The probabilities of misclassification tend to be underestimated in discriminant analysis. Bias can be estimated by a leave-one-out strategy or the bootstrap method (Efron, B. 1979. Bootstrap methods: another look at the jackknife. Annuls of Statistics 7(1):1-26).

A SYNTHETIC APPROACH TO PRINCIPAL COMPONENT ANALYSIS OF BIRD/HABITAT RELATIONSHIPS[1]

John T. Rotenberry[2] and John A. Wiens[3]

Abstract.--The application of principal components analysis (PCA) to bird/habitat relationships has essentially followed two paths: 1) species are ordinated based on PCA of average habitat values for individuals; 2) plots ordinated based on their average habitat values, and species' abundances on those plots correlated with the resulting component axes (which presumably reflect underlying environmental gradients).

Our proposed synthetic method is plot-based, but requires that sample points within plots be classified as lying within or outside of each individual species' area of use. As in (2), the total environmental variation, or multidemensional "habitat space", is defined by PCA of plot habitat values. However, rather than subjecting habitat values for each species to an independent PCA as in (1), a simple methodology may be used to map each species in the habitat space described by the plot PCA.

Several advantages accrue to mapping species and plots in the same environmental space. By graphing contours of species densities in this multidimensional space, patterns of abundance/habitat relationships that are not apparent from simple correlational analysis may emerge. Comparisons of plot means with values for individual species within a plot may reveal active habitat selection, or even consistent patterns of within-plot habitat partitioning between two species. Comparisons of density contours may suggest the presence of biological interactions, such as competition or ecological replacement, between two or more species.

The use of this technique is illustrated by analysis of 22 structural habitat variables collected at 26 North American grassland and shrubsteppe sites.

Key words: Birds; density contours; gradient analysis; principal components analysis; shrubsteppe; vegetation structure.

---

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Research Associate, Shrubsteppe Habitat Investigation Team, Department of Biology, University of New Mexico, Albuquerque, NM 87131. Present address: Assistant Professor, Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403.

[3]Professor, Shrubsteppe Habitat Investigation Team, Department of Biology, University of New Mexico, Albuquerque, NM 87131.

INTRODUCTION

The use of principal components analysis (PCA) in describing patterns of avian habitat occupancy has been most fruitful, both for consideration of the adaptations of individual species (e.g., James 1971) and analysis of community composition (Rotenberry and Wiens 1980). Indeed, PCA is now routinely performed as a matter of course on a wide variety of habitat parameters, not only for birds, but also for organisms spanning a large taxonomic range [e.g., Miracle 1974 (plankton), Johnson 1977a, b (bog plants)].

The application of PCA to bird/habitat relationships has essentially followed two paths that differ, to a certain degree, in their conceptual orientations. We propose a different, somewhat synthetic approach to PCA of avian habitats that combines the virtues of both previous approaches, yet retains a relatively simple and straightforward methodology. Results from the methodology are compatible with the concept of gradient analysis (Whittaker 1967, Terborgh 1971), and can be further interpreted in such a context.

RATIONALE AND METHODS

Previous Approaches

Bird Mean Habitat Vector

The first approach to bird habitat analysis can be called the bird mean habitat vector method (James 1971, Anderson and Shugart 1974, Whitmore 1975). In this method a sampling point is usually determined by the presence of a singing male bird. Once a point is located, a variety of habitat variables, generally assumed to be of ecological relevance, are measured in the immediate vicinity. Samples are taken for a number of different species at a number of different points, the average value of each variable is calculated for each species, and this set of averages then determines the mean habitat vector. These vectors are combined into $\underline{Y}$, the matrix of standardized variables for all species, and this matrix is analyzed using standard principal component techniques (e.g., Barr et al. 1976). Resulting orthogonal components are interpreted in light of their factor loadings (the correlation between new components and original variables). Habitat relationships among species are then reconstructed by plotting the location of each species in the newly defined component space, using their factor scores as coordinates. These scores for bird species are given by

$$\underline{F}_b = \underline{S}_b' \, \underline{R}_b^{-1} \, \underline{Y}_b', \qquad (1)$$

where $\underline{R}$ is the correlation matrix of the original variables, and $\underline{S}$ is the factor structure matrix containing factor loadings (notation based on that of Thorndike 1978). The subscript b denotes bird species data.

While such a technique clearly has considerable heuristic value (as, for example, in describing relative positions of species in multivariate habitat space), it also has two potential shortcomings: 1) by focusing mainly on the simple presence of a species, the advantages of a community approach are sacrificed, and observations concerning species diversity, resource partitioning, or relative widths of ecological habitat niches are not possible; and 2) because data are collected only on the basis of a species' presence, there is no informataion concerning variation in the species' numerical abundance as habitat changes.

Site Mean Habitat Vector

The usual alternative approach is in fact community oriented, although not without drawbacks of its own (e.g., Cody 1975, Rotenberry and Wiens 1980). The method begins with selection of a large series of plots or transects, generally chosen to be representative of variation in either some set of habitat variables or some set of bird species. Habitat variables of interest are measured at a number of randomly selected points on each plot and their average taken to yield what may now be called the mean habitat vector for the site. All plots are combined in $\underline{Z}$, the set of standardized variables for plots, and PCA is performed. The factor structure matrix again provides interpretation of the components and sites are ordered in this multidimensional space by their factor scores, given by

$$\underline{F}_s = \underline{S}_s' \, \underline{R}_s^{-1} \, \underline{Z}_s' \qquad (2)$$

where $\underline{S}$ and $\underline{R}$ are the factor structure matrix and correlation matrix for site habitat data (denoted by the s subscript).

Because these are plot-based samples, each site also has associated with it species abundances, diversities, or any other attribute of the avian community one cares to calculate, and these can then be correlated with site factor scores. Significant individual correlations are generally interpreted as representing species' responses to the multidimensional habitat spectrum represented by the site ordination. If species diversity seems to vary in some meaningful manner after the sites have been ordered, then one can speculate about the nature of community organization along the gradient. These sorts of interpretations are not possible under the bird mean habitat vector method, but unfortunately their gain is offset by loss of some information about individual species. Because a species cannot be considered separately from a site, the point for that site represents all species simultaneously. If there is any within-plot habitat selection or partitioning by species, for example, this will be completely obscured, and any generalizations that one might want to make about species' relationships could be compromised.

Synthetic Approach

The technique we propose combines what we think are useful elements of both site and species ordinations. Although the methodology is largely based on site-oriented sampling, slight modifications of traditional methodology yields data that are compatible with the species-oriented approach as well.

As in the site-oriented approach, a series of sites are selected that encompass some environmental range in which one is interested, and attributes of both habitat and bird populations on these sites are measured. In the course of estimating bird densities, it is generally possible to estimate microhabitat or within-plot use by the individuals of a species as well. If, for example, individuals are territorial, this becomes a simple exercise in mapping territorial boundaries. Superimposed upon this are locations of the random points at which habitat variables are measured. These points may be characterized as lying within or outside of the area used by a species (e.g., Wiens 1969), and those that lie within the use-area can be used to create a mean habitat vector for that species at that site. These vectors will likely differ for different species at the same site (depending on the degree of within-plot spatial or habitat overlap between them), or for the same species at different sites. To the extent that a species is nonrandomly selecting habitat within the plot, its vector will differ from that of the plot as a whole. The plot or site mean vector, of course, is determined by all points.

The multidimensional environmental space in which all samples have been taken is defined by a PCA of the matrix of standardized variables for the sites, $\underline{Z}$, and as such does not differ from that presented above (equation 2). PCA is still picking out the major patterns of covariation in habitat variables that are latent within our overall selection of plots. One elects, of course, to concentrate on those components that are relatively strong (as evidenced by the relative magnitude of their eigenvalues) and meaningful, in that the patterns of their factor loadings appear to make some sort of ecological sense. Because habitat variables are in fact environmental measures, the components can be interpreted as representing real-world ecological gradients, and one can begin to apply concepts of gradient analysis to the sites, and now to the species as well.

The next step is to plot species along the same gradients. To do this, new factor scores are calculated using the factor structure matrix and correlation matrix from the habitat variables at sites, and the matrix of standardized variables from birds. Thus,

$$\underline{F}_b = \underline{S}_s' \, \underline{R}_s^{-1} \, \underline{Y}_b', \qquad (3)$$

where all matrices are as in equations (1) and (2). Resulting factor scores thus map the habitat selection of birds ($\underline{Y}_b$) onto the multidimensional environmental gradients defined by sites ($\underline{S}_s$ and $\underline{R}_s$). In other words, sites are used regardless of their species composition to determine the habitat patterns, then the distribution of the birds is plotted along these patterns, regardless of the sites on which they occur.

There are several advantages to this methodology. First is that as data are collectd from relatively large plots or transects, the community-oriented attributes that are lacking in the species-specific approach are retained, as well as estimates of individual species abundances. At the same time, however, single species may be ordinated or otherwise related to one another on the basis of species-specific, not site-specific, responses to the multivariate habitat gradient. The second major advantage is that although species at a site are analyzed in the same environmental space as the site, they need not be equated with that site's mean values. This means, for example, that within-plot habitat selection will not be obscured (fig. 1). Under normal site-based analysis, the point that represents the plot also represents, in this case, three species. By this synthetic method, however, if a species occupies a distinctly different subset of habitat than the average for a site on which it occurs, such will be readily apparent by the degree of departure of the species' position from the site's position in the PCA-space (fig. 1). Because the point for the site represents all species, any sort of within-plot habitat partitioning between them that may be occurring will be obscured using the site mean habitat vector approach. Such partitioning may be detected under the synthetic methodology, however, if there are consistent patterns of displacement of two species that otherwise co-occur at a number of sites (e.g., species 1 and species 2 in fig. 1).

Perhaps one of the most interesting properties of site-oriented analyses arises from the fact that each site-specific point for a species can be associated with that species' density at the site as well. By plotting these densities on the derived component axes, one can begin to build a picture of the quantitative distribution of species with respect to environmental gradients represented by these axes. If there are a sufficient number of sample points, contours of a species' abundance patterns may be plotted as well. The contours shown in figure 2 present a considerably stylized example, but they can be used to demonstrate some patterns that might arise from an analysis of this sort. This set of contours, for example, shows this species' numerical response to the derived habitat gradients, a response that is unlikely to be detected through any correlational analysis because of its intrinsic nonlinearity. Further, projection of contours onto each of the axes separately indicates that this species is rather

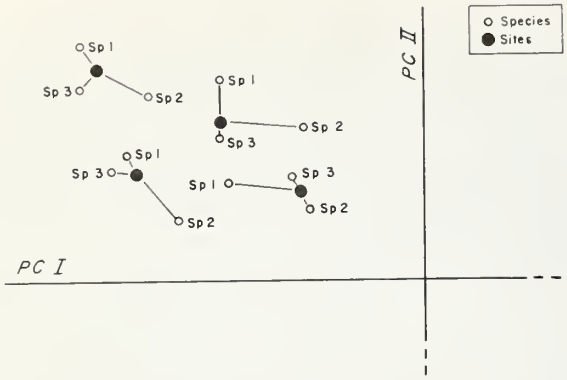


Figure 1. Hypothetical species and sites plotted in environmental space defined by the first two principal components (PCI and PCII) of site-based environmental variables. Lines connect species to sites on which they occurred.

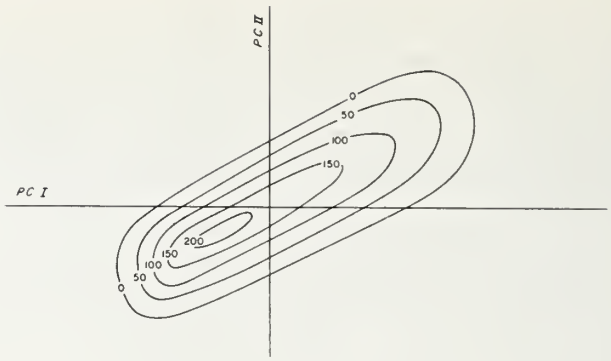


Figure 2. Hypothetical contours of species abundance patterns plotted in environmental space defined by the first two principal components (PCI and PCII) of site-based environmental measures. Contours represent isopleths of density (individuals/km^2).

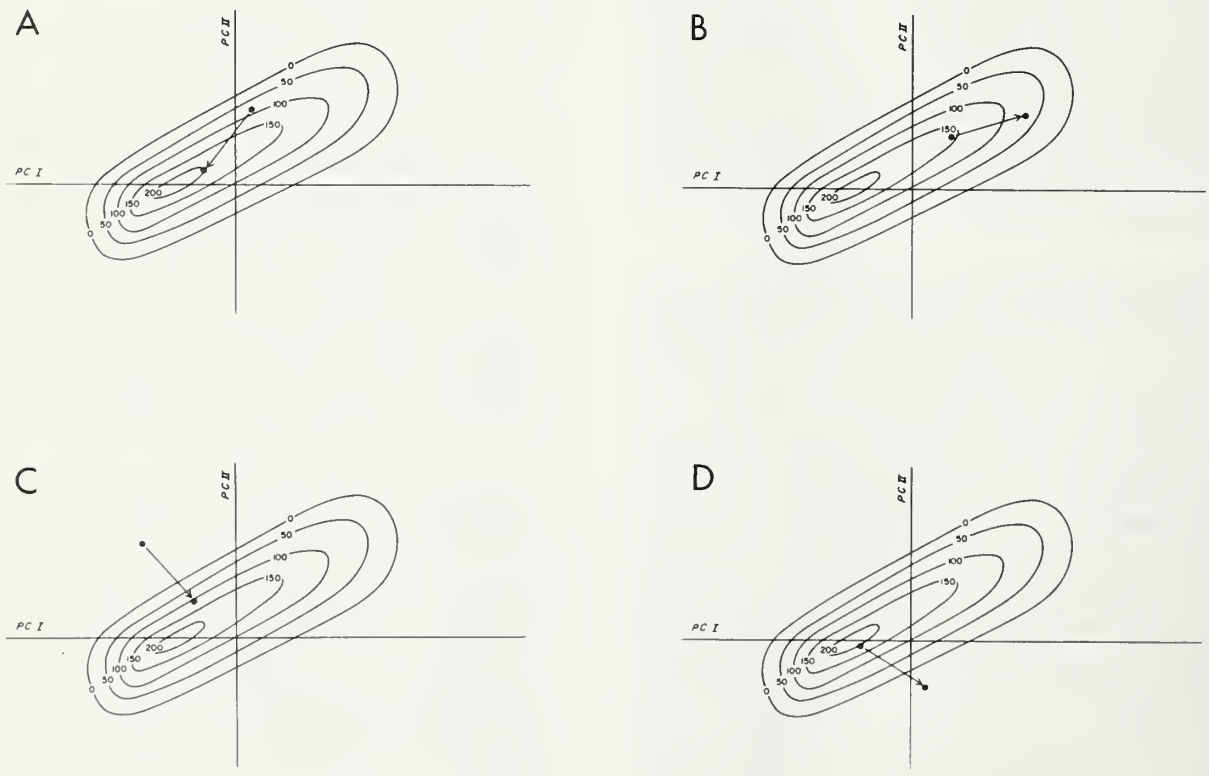


Figure 3. Hypothetical contours of species abundance patterns. Axes and isopleths as in figure 2. Arrows denote change in site characteristics as a result of habitat alteration. Changes in site characteristics may effect the following changes in a species' abundance at the site: A-increase; B-decrease; C-local invasion; D-local extinction.

generalized with respect to its distribution along these gradients. Although still maintaining the same basic configuration of contours, a different species, on the other hand, might occupy a relatively wide range on one axis but a much narrower range on the other--we can thus identify what may be called axes of specialization and generalization.

Perhaps the most interesting aspect of these contours is their potential in habitat management predictions. Insofar as one has some idea of how an environmental alteration will affect a site's location along each of the gradients--that is to say, to the degree that one can predict how a site will "move" in multidimensional habitat space after some sort of treatment--then one should be able to estimate the effect of that treatment on a species' population. For example, any sort of alteration that caused a plot to move in habitat space as indicated in figure 3A would likely result in an increase in abundance of this hypothetical species, while a different change (fig. 3B) may be much more likely to result in a decrease. We might even be able to predict changes in habiat that would lead to invasion of a species into the area (fig. 3C), assuming, of course, that it were biogeographically feasible for it to do so. Perhaps most important from a management standpoint, one might be able to define a habitat alteration that would lead to a species' local extinction (fig. 3D).

The sorts of contours depicted in figure 3 are, of course, stylized, and more often than not those derived from sets of real data are likely to deviate markedly from such smooth patterns. Certainly one of the biggest contributors to an uneven species distribution will be uneven sampling intensity with respect to the derived gradients; unfortunately such omissions are apparent only after the fact and are thus difficult to control. Other reasons, however, are more biological in nature. For example, one may sample at the periphery of habitat that is suitable for a species and thus may map only a portion of its contours (fig. 4A). Alternatively, a species may be distributed in a non-Gaussian fashion along both habitat gradients (cf. Colwell and Futuyma 1971), which would yield a pattern similar to that of figure 4B. The most extreme expression of this type of response would be a species' recognition of an environmental discontinuity or ecotone on what are otherwise statistically continuous gradients. Such an ecotonal response would likely be evidenced by a very rapid change in a species' abundance over an apparently short environmental distance (fig. 4C). On the other hand, such a pattern might also correspond with sharp abutment of one species upon another (fig. 4D). Although such a pattern may still reflect an ecotonal response, some sort of biological interaction between the two species, such as competition, becomes more likely. This pattern also identifies habitat in which investigations of such interactions should be conducted. Experiments or other comparative analyses conducted in such areas identified by this technique are likely to be more rewarding than those conducted in locales randomly chosen throughout the species' ranges.

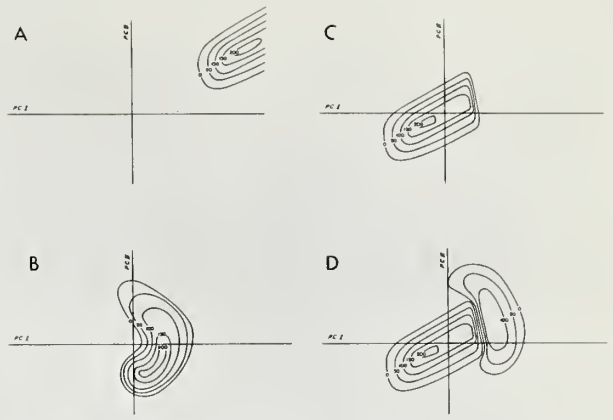


Figure 4. Hypothetical deviation from the "normal" species contours of figures 2 and 3. Axes and isopleths as in figure 2. A-Species peripheral to habitat samples. B-Non-Gaussian distribution along both environmental gradients. C-Species perceives and responds to "ecotone" on otherwise statistically continuous gradient. D-Competitive interaction between species results in habitat displacement.

If one is inclined to treat the derived gradients as representing habitat niche axes of one sort or another [and there seems to be ample reason to do this for birds (Rotenberry, 1981)], then there arises yet another set of calculations that can be made. Niche width, for example, can be estimated from some measure of dispersion around the species' centroid in multivariate space. Niche overlap between species may also be calculated, incorporating both distances between two centroids and changes in the species' contours as well. While we offer no speculation on the specific form such breadth and overlap statistics might take, the distributions of species in this environmental space will likely, we think, provide a fertile field for the application of many traditional (and even nontraditional) niche metric manipulations.

AN EXAMPLE: NORTH AMERICAN STEPPE AVIFAUNA

We would now like to provide a brief demonstration of the application of this methodology and some of its concepts to a real set of data, although unfortunately not one originally collected explicitly for this purpose. The data come from a collection of 26 grassland and shrubsteppe sites scattered throughout middle and western North America, representing a wide array of steppe vegetation types (fig. 5). Sampling methodology, site locations, species lists and abundances, descriptions of variables, etc. are fully detailed elsewhere (Rotenberry and Wiens 1980). At each plot we set up a 10-ha grid, mapped the territories of all breeding birds, and measured habitat variables. In this particular

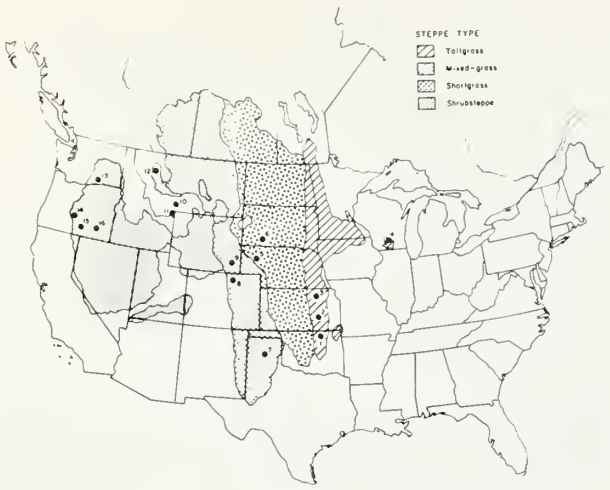


Figure 5. Location of avian censuses used in this analysis. Numerals keyed to table 1 in Rotenberry and Wiens (1980). Steppe types generalized from Küchler (1964).

study we were interested in the role of spatial heterogeneity, or patchiness in vegetaion structure, in determining distribution and abundance of grassland birds and the structure of their communities. The 22 variables that were measured fell into two basic categories: coverage variables (simply the percent coverage of various physiognomic classes, such as shrubs, grasses, forbs, or bare ground) and structural variables. Structural variables are also physiognomic in nature, but have the additional property of "dimension"; that is to say, variation in a structural measure is generally associated with variation in either a horizontal or a vertical plane. Ultimately 10 horizontal heterogeneity indices, 5 vertical indices, and 7 coverage classes were considered. The results, given here in the most basic form (table 1), were somewhat surprising in that a very distinct separation between the vertical and horizontal indices were observed; 9 of 10 horizontal indices loaded high on the first axis (41% of total variation), while all 5 vertical indices loaded high on the second (22% variation). Increasing horizontal heterogeneity was associated with increasing coverage of shrubs and bare ground, and decreasing coverage of grass and litter. Vertical heterogeneity varied slightly with changing forb coverage. Together, these two axes accounted for almost two-thirds of the total variation. We thus interpreted these components as representing two largely independent gradients in vegetation structural heterogeneity--a major axis of horizontal patchiness and a minor axis in vertical patchiness.

The sites tended to sort themselves into broad classes along these axes in horizontal and vertical heterogeneity (fig. 6A). Tallgrass sites evidenced substantial vertical heterogeneity, but relatively little patchiness in a horizontal plane. Shrubsteppe sites generally showed as much

Table 1. Factor loadings of vegetation principal components. Only those loadings greater than 0.40 (analogous to a correlation significant at the 0.05 probability level) are shown; only the first two of five factors with eigenvalues greater than 1 are shown. Complete details of the analysis, including explanations of variables, are given by Rotenberry and Wiens (1980).

	Factor:	I	II
	Eigenvalue:	9.02	4.94
Vegetation variable	% σ:	41.0	22.4
	Σ % σ:	41.0	63.4
Structural-horizontal			
HIT-10		0.46	0.44
LITDEP		-0.71	0.52
CVTOTHIT		0.83	
CVMAXHGT		0.84	
CVLITDEP		0.80	
CVTOTHGT		0.50	
CVHGTDIF		0.68	
HITS-HI		0.93	
LIT-HI		0.69	
DIST-HI			0.73
Structural-vertical			
TOTHITS		0.56	0.56
MAXHGT			0.77
EFFHGT			0.86
TOTHGT			0.88
HGT-HI			0.70
Coverage			
Grass		-0.87	
Forb			0.40
Shrub		0.80	
Bare		0.92	
Litter		-0.82	

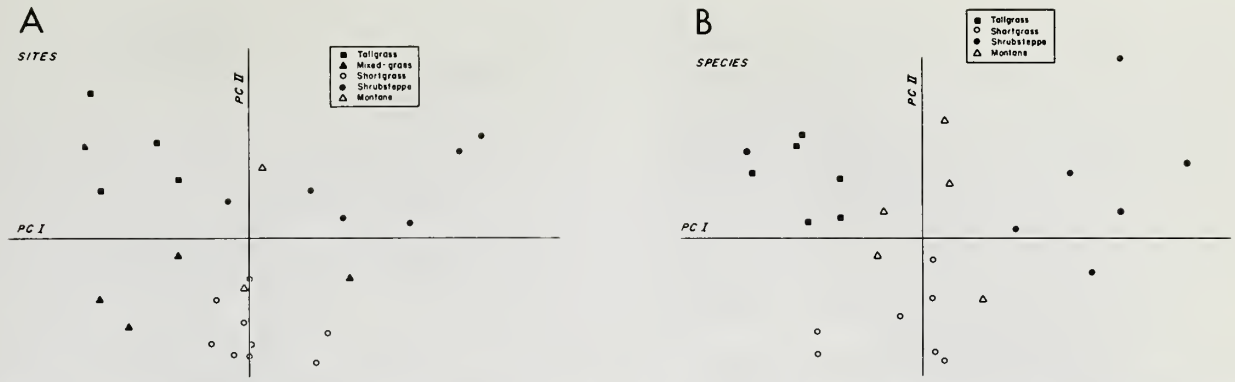


Figure 6. Distribution of steppe bird census sites (A) and bird species means (B) in environmental space defined by the first two principal components (PCI and PCII) of 22 site-based habitat structure variables. PCI represents increasing horizontal heterogeneity; PCII represents increasing vertical heterogeneity.

vertical heterogeneity as the tallgrass ones, but differed markedly in being much patchier horizontally. Shortgrass sites were intermediate in horizontal heterogeneity but, as one might expect, showed very little vertical variability. Mixed-grass sites were the most varied in their structure, with some positioned intermediate to shortgrass and tallgrass plots, while another more closely resembled shrubsteppe sites in structure. Montane sites were intermediate as well.

Figure 6B shows bird species plotted in the same space; each point represents the average of a species' factor scores for all of its occurrences combined without regard to its abundance. Not surprisingly, species that are identifiable as being largely tallgrass, shortgrass, or shrubsteppe species are generally found in the same areas of PCA-space as their sites. Of more interest, of course, are some details that go into generating these means. For example, figure 7 reprsents both sites and birds for four tallgrass and four shrubsteppe samples. (Scientific names of all species are given in appendix I.) It is apparent that a number of shrubsteppe species are not very close to their site means, and indeed some sites appear to have a great deal of dispersion around them. Although tallgrass species did not appear as dispersed, dickcissels and grasshopper sparrows did demonstrate a consistent pattern of within-plot partitioning, with the latter found in vegetation of less vertical heterogeneity but greater horizontal patchiness. The co-occurring eastern meadowlark, however, seemed to be positioned independently.

Contour intervals for the species show a wide variety of patterns (figs. 8-15). Our collection of sites appears to have almost exactly centered on the habitat distribution of western meadowlarks (fig. 8), and this pattern, in retrospect, is perfectly consistent with most of what is known about the behavior, habitat selection, and

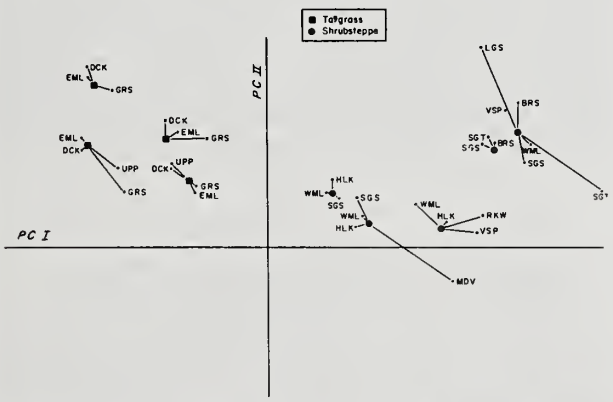


Figure 7. Distribution of tallgrass and shrubsteppe bird census sites and individual bird species from those sites. Axes as in figure 6. EML = eastern meadowlark, DCK = dickcissel, GRS = grasshopper sparrow, UPP = upland plover, WML = western meadowlark, HLK = horned lark, SGS = sage sparrow, MDV = mourning dove, RKW = rock wren, VSP = vesper sparrow, SGT = sage thrasher, BRS = Brewer's sparrow, LGS = loggerhead shrike.

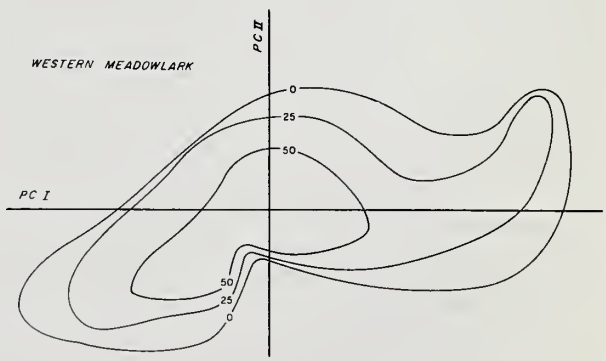


Figure 8. Distributional pattern of western meadowlarks. Axes as in figure 6. Isopleths as individuals/km².

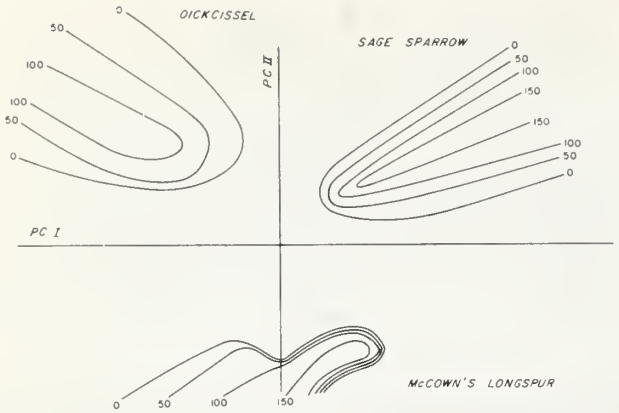


Figure 9. Distributional pattern of dickcissels, sage sparrows, and McCown's longspurs. Axes and isopleths as in figure 8.

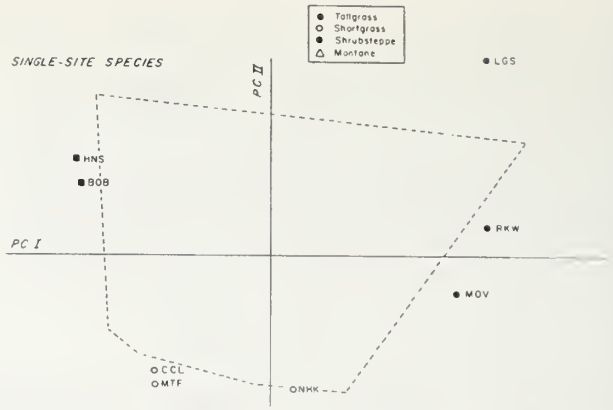


Figure 10. Distributional pattern of species that occurred at only a single census site. Species associated with montaine meadows have been omitted. Axes as in figure 6. Dashed line encloses all site mean values (figure 6A). HNS = Henslow's sparrow, BOB = bobolink, CGL = chestnut-collared longspur, MTP = mountain plover, NHK = common nighthawk, MDV = mourning dove, RKW = rock wren, LGS = loggerhead shrike.

geographical distribution of this species (Lanyon 1956, and personal observation).

It also appears that the range of habitats sample was only peripherally used by several species. For example, the dickcissel (a tallgrass prairie species), sage sparrow (a shrubsteppe bird), and McCown's longspur (abundant in shortgrass) all seem to have centers of distribution that lie at the edge or even outside of the limits of habitat gradients defined here (fig. 9). Presumably, then, if data were collected in more extreme shrubsteppe vegetation, we would be able to define the limits of sage sparrow distribution more accurately. The most extreme example of peripheral species are those that occurred on just one of the 26 plots we sampled (fig. 10). If a line is graphed that encloses all the site mean values for these gradients, it is apparent that these species all lie beyond that boundary, indicating that even at those sites on which they did occur they were occupying peripheral habitat. It likewise is assumed here that if sampling effort were increased away from this boundary, more sites at which these species occur would be encountered.

Horned larks, present in both shrubsteppe and shortgrass habitats, apparently increase in abundance as the habitat becomes more uniform in both dimensions (fig. 11). This is consistent with our own observations of the species, which suggest that the lark's ultimate idea of habitat nirvana is the uniform monotony of a paved parking lot.

Some patterns of contours are more difficult to interpret. Brewer's sparrows, for example, evidenced a rather discontinuous distribution (fig. 12). While we have searched for another species that might fit in the gap, none of the species enounterd in our censuses througout the North American steppe vegetation showed a reciprocal distribution in this area. At this time we can only surmise that there may be some sort of biogeographical constraint to Brewer's sparrows filling in the vacancy, or that perhaps

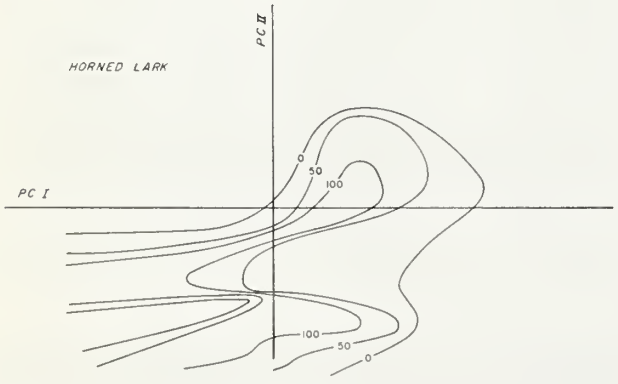


Figure 11. Distributional pattern of horned larks. Axes and isopleths as in figure 8.

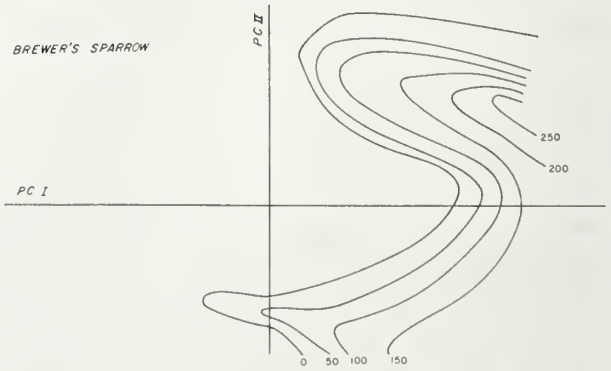


Figure 12. Distributional pattern of Brewer's sparrows. Axes and isopleths as in figure 8.

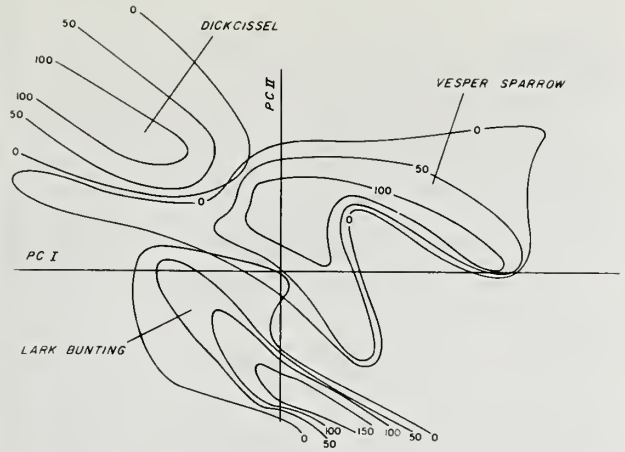


Figure 13. Distributional pattern of vesper sparrows, dickcissels, and lark buntings. Axes and isopleths as in figure 8.

populations in the two rather disjunct habitat types represent different ecological races or even incipient subspecies.

The precise habitat affinities of vesper sparrows have defied ready generalizations, and their distribution along these synthetic habitat gradients (fig. 13), when considered by itself, adds little to a more precise definition. It is interesting, however, to plot the vesper sparrow simultaneously with the dickcissel (a tallgrass species) and the lark bunting (a shortgrass species) (fig. 13). While one cannot conclude from this analysis what mechanism might ultimately be responsible for these observed distributional patterns, they certainly appear to be nonrandom.

One may also examine situations where nonrandom patterns of distribution might be predicted a priori, such as might result from competitive interactions. Eastern and western meadowlarks, for example, are congeneric and are thought to express a number of competition-induced relationships. Their contours, while largely exclusive, do show some overlap (fig. 14). Although no really clear pattern is present in terms of a sharp abutment or exclusivity, we could predict that sites that fall within the particular area where density overlap is greatest are where interspecific territoriality is likely to be most intense. Although dickcissels and grasshopper sparrows appeared to partition individual tallgrass plots (fig. 7), their pattern of contours (fig. 15) suggests little partitioning on a regional scale and instead seems indicative of independent differential habitat selection.

Finally, one can examine the predicted effect of some major alteration in some part of the steppe environment; say, for example, a decision were made to mow a tallgrass prairie? One expects a tremendous drop in vertical heterogeneity as the grasses are all cut flat, with perhaps a slight increase in horizontal patchiness as new areas of bare ground open up following removal of forbs and small shrubs. By superimposing this change upon the species' contours, changes in species abundances can be predicted (fig. 16). Clearly the tallgrass species such as dickcissels (fig. 16A) and eastern meadowlarks (fig. 16B) are likely to disappear altogether, while species abundant in habitat with low vertical heterogeneity, like horned larks (fig. 16C) or lark buntings (fig. 16D), will probably increase considerably if there are no biogeographical constraints to their response to this habitat alteration. Grasshopper sparrow densities may change (fig. 16E), but it seems unlikely that the species would disappear altogether. In the case of the western meadowlark, however, predictions could be equivocal (fig. 16F); this will depend both upon accuracy in estimating how an altered site will "move" in habitat space and upon proximity of the new site position to contours representing an area of fairly rapid change in a species' abundance.

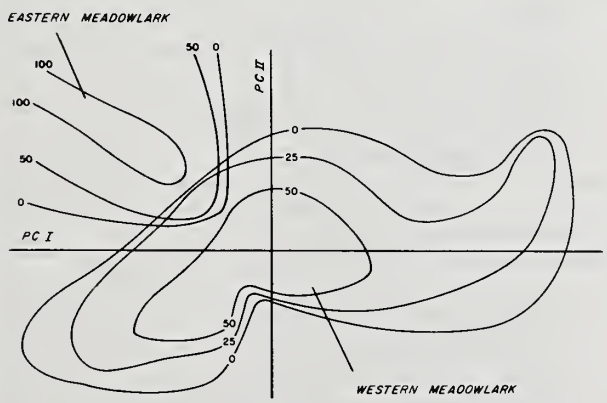


Figure 14. Distributional pattern of eastern and western meadowlarks. Axes and isopleths as in figure 8.

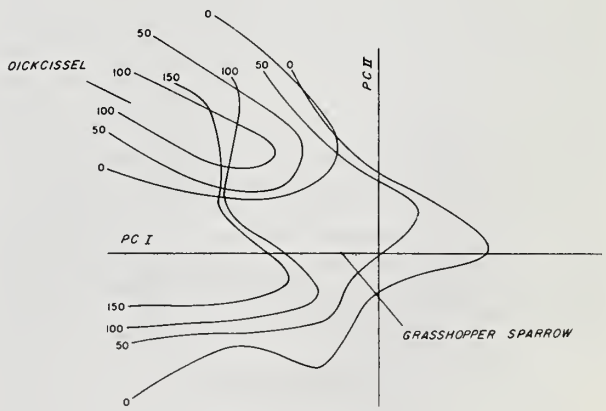


Figure 15. Distributional pattern of dickcissels and grasshopper sparrows. Axes and isopleths as in figure 8.

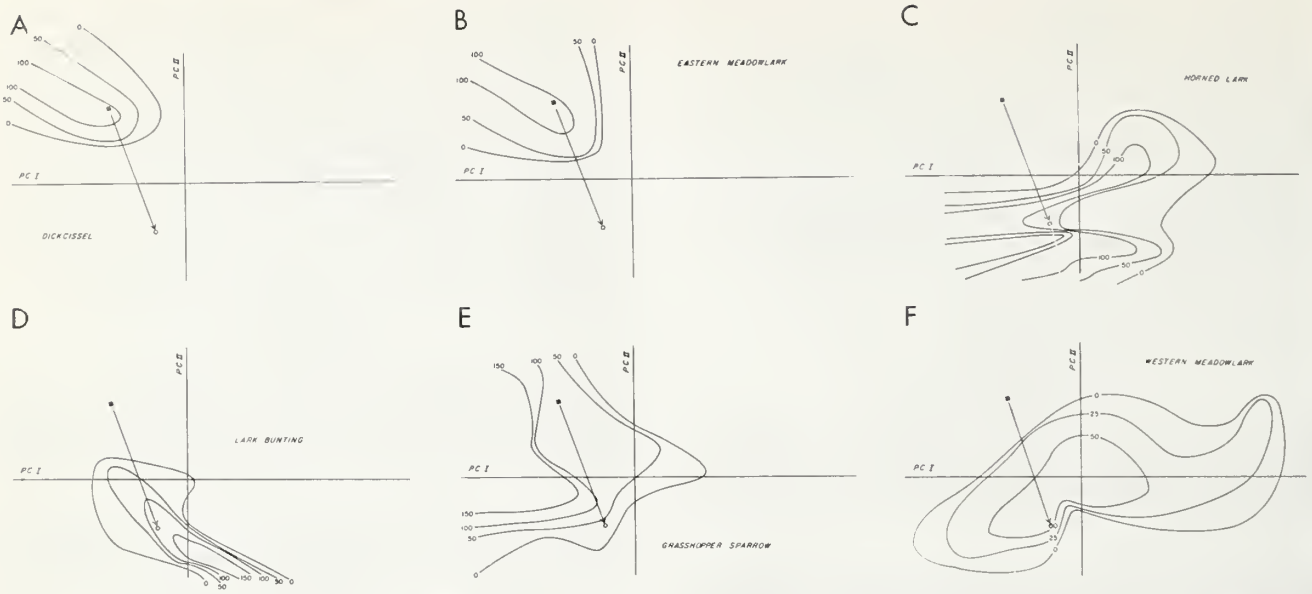


Figure 16. Responses of affected species to hypothetical structural changes and associated multidimensional movement of a tallgrass site that is subjected to mowing. A-dickcissel, B-eastern meadowlark, C-horned lark, D-lark bunting, E-grasshopper sparrow, F-western meadowlark.

CONCLUSIONS

We hope to have made it clear that, even with a set of data that was not taken with the methodology in mind, our technique of combining the individual species approach with one that is plot-oriented is feasible; indeed, there may be sets of existing bird/habitat data to which it may be applied. Analysis certainly need not be confined to physiognomic variables as we have done here, but can be extended to whatever environmental or habitat variation one thinks is important in affecting species distributions and relationships. Although certain biological objections can be raised regarding application of PCA to either species- or plot-oriented data (Johnson 1981), we believe this new technique, combined with judicious selection of variables, appropriate sample sizes, and other attention to statistical niceties, may ameliorate these objections substantially. One must bear in mind that the technique of PCA is descriptive rather than inferential and that it basicially provides us a sophisticated "multivariate natural history". We think that such descriptions will ultimately prove robust with respect to what is already known about the particular set of species described and that they will be capable of generating new insights for those that employ them.

ACKNOWLEDGMENTS

Field data were gathered during studies supported by National Science Foundation grants GB-6606 to J.A. Wiens and GB-7824, GB-13096, GB 31862x, GB-31862, and DEB73-02027 A03 to the Grassland Biome, United States International Biological Program. The final stages of the research were supported by NSF grant BMS 75-11898. Data analysis was supported by the University of New Mexico Computer Center.

LITERATURE CITED

Anderson, S.H., and H.H. Shugart. 1974. Habitat selection of breeding birds in an east Tennessee deciduous forest. Ecology 55:828-837.

Barr, A.J., J.H. Goodnight, J.P. Sall, and J.T. Helwig. 1976. A user's guide to SAS76. SAS Institute, Inc., Raleigh, N.C.

Cody, M.L. 1975. Towards a theory of continental species diversities. p. 214-157 In Cody, M.L., and J.M. Diamond, editors. Ecology and evaluation of communities. Belknap Press, Cambridge, Mass.

Colwell, R.K., and D.J. Futuyma. 1971. On the measurement of niche breadth and overlap. Ecology 52:567-576.

James, F.C. 1971. Ordinations of habitat relations among breeding birds. Wilson Bulletin 83:215-236.

Johnson, D.H. 1981. How to measure habitat--a statistical perspective. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Johnson, E.A. 1977a. A multivariate analysis of the niches of plant populations in raised bogs. I. Niche dimensions. Canadian Journal of Botany 55:1201-1210.

Johnson, E.A. 1977b. A multivariate analysis of the niches of plant populations in raised bogs. II. Niche width and overlap. Canadian Journal of Botany 55:1211-1220.

Küchler, A.W. 1964. Potential natural vegetation of the conterminous United States. American Geographical Society Special Publication No. 36, New York, N.Y.

Lanyon, W.E. 1956. Ecological aspects of the sympatric distribution of meadowlarks in the north-central states. Ecology 37:98-108.

Miracle, M.R. 1974. Niche structure in zooplankton: a principal components approach. Ecology 55:1306-1316.

Rotenberry, J.T. 1981. Why measure bird habitat? In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Rotenberry, J.T., and J.A. Wiens. 1980. Habitat structure, patchiness, and avian communities in North American steppe vegetation: a multivariate analysis. Ecology 61:1228-1250.

Terborgh, J. 1971. Distribution on environmental gradients: theory and preliminary interpretation of distributional patterns in the avifauna of the Cordillera Vilcabamba, Peru. Ecology 52:23-40.

Thorndike, R.M. 1978. Correlation procedures for research. 340 p. Gardner Press, New York, N.Y.

Whitmore, R.C. 1975. Habitat ordination of passerine birds in the Virgin River valley, southwestern Utah. Wilson Bulletin 87:65-74.

Whittaker, R.M. 1967. Gradient analysis of vegetation. Biological Review 42:207-264.

Wiens, J.A. 1969. An approach to the study of ecological relationships among grassland birds. 93 p. Ornithological Monographs 8.

Appendix I. Scientific names of all birds mentioned in text or figures.

Mountain plover	Eupoda montana
Upland plover	Bartramia longicauda
Mourning dove	Zenaida macroura
Common nighthawk	Chordelis minor
Horned lark	Eremophila alpestris
Rock wren	Salpinctes obsoletus
Sage thrasher	Oreoscoptes montanus
Loggerhead shrike	Lanius ludovicianus
Bobolink	Dolichonyx oryzivorus
Eastern meadowlark	Sturnella magna
Western meadowlark	Sturnella neglecta
Dickcissel	Spiza americana
Grasshopper sparrow	Ammodramus savannarum
Henslow's sparrow	Passerherbulus henslowii
Lark bunting	Calamospiza melanocorys
Vesper sparrow	Pooecetes gramineus
Sage sparrow	Amphispiza belli
Brewer's sparrow	Spizella breweri
McCown's longspur	Calcarius mccowni
Chestnut-collared longspur	Calcarius ornatus

DISCUSSION

BOB WHITMORE: How did you compute the density contours? The smooth curves that you presented would seem to indicate many ($>10^8$) data points. How many points per species did you have?

JOHN ROTENBERRY: Not nearly as many as we would have liked, especially for species whose distributions take them off the edges of the habitat space. The curves were initially drawn using strict linear interpolation between points; I subsequently smoothed them by eye for this presentation. To a considerable extent it represents artistic license tempered by the interpolated realities and biological intuition.

BOB WHITMORE: You had some species with "peripheral ranges" on your ordinations. How do you know that the observed distributions are not just an artifact of the variables being measured?

JOHN ROTENBERRY: We cannot know for sure, but it seems likely due to biogeographical considerations and the distribution of our sites.

E. JAMES HARNER: How did you scale the variables in your PCA analysis? Did you only use the correlation matrix?

JOHN ROTENBERRY: The variables were all normalized using either log or arcsin square root transformation. I only used the correlation matrix for PCA, which is analogous to using the covariance matrix of standardized data. The component axes were also rotated using varimax.

BOB WHITMORE: Given the overlap problems when using centroid distances, why not use a multivariate measure of niche overlap?

JOHN ROTENBERRY: I think that would be the best approach.

JAMES DUNN: How the species counts balance each other as a function of site properties is an interesting question. Does your analysis reflect this? Why not try the GSK model as implemented by PROC FUNCAT(SAS)? One nice feature of the model is that it would allow formal tests of the effects of habitat variation on relative frequencies of the species. In particular, the total bird count/plot could be used as a predictor to see if site productivity affects the relative composition of the species. The only difficulty I see with the FUNCAT is zero counts which you may have. Potentially a maximum likelihood solution could handle that using "working values."

JOHN ROTENBERRY: In the paper where the data were originally described, we did test correlations of species diversity, richness, and evenness with each of the component axes. Diversity did significantly increase with incoming vertical heterogeneity, but varied independently of horizontal heterogeneity. Richness and evenness were not significantly correlated with either axes. With respect to examining relative frequencies, I think it might be worthwhile, but we are, in fact, plagued with zero counts throughout the bird data set. I'm not familiar with FUNCAT, so I don't know how the use of "working values" will affect the outcome.

CHARLES SMITH: When making measurements in the field, how does one differentiate between the extremes of the "niche-habitat continuum," that is, are habitat variables distinguished from niche variables when measurements are taken?

JOHN ROTENBERRY: I suspect that in general we are not dealing with the extremes, but that instead most of the variables we chose to measure reflected both aspects. My feeling is that the synthetic component axes do represent habitat gradients, and that (for birds at least) habitat is a niche dimension, for the reasons I suggested earlier in this workshop.

PAUL GEISSLER: Looking at your figures, I noted that many of the frequency ellipses were oriented at about 45° to the axis. Perhaps the interpretability of the axes for the birds might be improved if the axes were rotated parallel to the orientation of the ellipses.

JOHN ROTENBERRY: There are two problems with doing that. First, if we rotate the axes with respect to the birds, we destroy the relationships among the variables and change the interpretation of the axes in some unknown way. This would not be the same thing as varimax or orthomax rotation. Second, the ellipses do represent the relations of the bird species to the derived axes. If a species shows a response to both horizontal and vertical heterogeneity, we expect frequency ellipses to be oriented just this way.

ROBUST PRINCIPAL COMPONENT AND DISCRIMINANT ANALYSIS OF TWO GRASSLAND BIRD SPECIES' HABITAT[1]

E. James Harner[2] and Robert C. Whitmore[3]

Abstract.--Outliers in multivariate data can have pronounced effects on the interpretations and conclusions of statistical analyses. However, data is rarely examined for outliers, due in part to the difficulty in discovering them if the dimensionality of the variable space is greater than two or three. We also do not believe that researchers are aware of the severity of the problem that outliers create.

Techniques are currently being developed which are robust to small departures from the assumed model and, in particular, to outliers. The robust methods summarized in this paper are simple extensions of maximum likelihood estimators of the mean vector and covariance matrix assuming an underlying normal distribution. In essence, they are weighted versions of maximum likelihood techniques in which weights are determined iteratively based upon the size of Mahalanobis distances. Principal components can be determined from the robust covariance or correlation matrix. For the discriminant problem, robust estimates are determined for each group and these are substituted into Fisher's linear discriminant function.

These techniques are applied to two sparrow species on which four habitat variables were measured. These data have some outliers but the data set would not be classified as being "bad." The principal component analyses for both species resulted in less emphasis being placed on the first component for the robust techniques as compared to the classical method. Fisher's linear discriminant model depended too heavily on several outlying values. In order to improve the stability of estimates and the accuracy of future classifications, Huber's robust discriminant analysis with a cutoff value of 1.97 was adjudged to be best.

Key words: Discriminant analysis; grassland birds; maximum likelihood estimation; principal components analysis; robustness.

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Associate Professor of Statistics, Department of Statistics and Computer Science and Associate Statistician, Agricultural and Forestry Experiment Station, West Virginia University, Morgantown, WV 26506

[3]Associate Professor of Wildlife Ecology, Division of Forestry, West Virginia University, Morgantown, WV 26506.

INTRODUCTION

Multivariate analyses often start with mean vector and covariance matrix estimates. In principal components the eigenvectors and eigenvalues are computed from the covariance matrix or from scaled versions of it. The mean vector and covariance matrix are calculated from each sample in discriminant analysis. Unfortunately, classical maximum likelihood estimators based on multivariate normal theory and least squares estimators are extremely sensitive to outlying observations.

During the past 20 years statisticians have been developing new estimators or modifying existing estimators to make them less sensitive to certain departures from the assumptions of a model. Three classes of estimators are being developed:

1) M-estimators or maximum likelihood type estimators;
2) R-estimators or rank type estimators; and
3) L-estimators or estimators based on linear combinations of order statistics.

Estimators of the above types are called robust estimators. By robustness we mean that the distribution of an estimator is insensitive against small deviations in the shape of the underlying distribution of the assumed model. This is essentially equivalent to an estimator being robust to outliers although other distributional variations are possible. In many biological sampling situations outliers frequently arise due to gross errors (bad data points) or from heavy-tailed distributions. Huber (1977a) states that "5-10% wrong values in a data set seem to be the rule rather than the exception." Since our experience with outliers in biological data is consistent with this statement, we feel robust methodologies are of great value to researchers.

A general theory of R and L-estimators is currently being developed for the univariate linear model (Hettmansperger and McKean 1977, Bickel 1973). However it is not clear how to extend these to the multivariate case. The theory of M-estimators was begun by Huber (1964). Huber (1977a, 1977b) and Maronna (1976) developed M-estimation techniques for the mean vector and covariance matrix of a single population. They derived properties of these estimators and suggested algorithms to compute them. Randles et al. (1978) studied the performance of these and other estimators in the two group linear and quadratic discriminant analysis problem. Other than these papers, little else has been done.

The intent of this paper is to apply the estimators developed by Huber and Maronna to bird habitat data. Both principal component and discriminant analyses are presented. First an introduction to M-estimators is given.

M-ESTIMATORS

Univariate Problem

For intuition the univariate location problem is considered first. Let $f(x-\mu)$ represent a density and μ the location parameter (mean) we want to estimate. $X_1, X_2, \ldots, X_n$ represents a random sample from this distribution. The maximum likelihood estimate is that value of μ which maximizes

$$L(\mu) = \prod_{i=1}^{n} f(x_i-\mu),$$

the likelihood function. It is more convenient to maximize $\ln L(\mu)$ which gives the same solution since the natural logarithm is a monotonic transformation. Thus we maximize

$$\ln L(\mu) = \sum_{i=1}^{n} \ln f(x_i-\mu) = -\sum_{i=1}^{n} \rho(x_i-\mu)$$

with respect to μ where $\rho(x-\mu) = -\ln f(x-\mu)$. Differentiating and setting the resulting expression to zero we get

$$d[\ln L(\mu)]/d\mu = -\sum_{i=1}^{n} f'(x_i-\mu)/f(x_i-\mu)$$

$$= \Sigma\phi(x_i-\mu) = 0,$$

where $\phi(x-\mu) = d[-\rho(x-\mu)]/d\mu$.

The maximum likelihood estimate, $\hat{\mu}$, solves the last equation.

The normal case is given by:

$$f(x-\mu) = (1/\sqrt{2\pi})e[-(x-\mu)^2/2],$$

$$\rho(x-\mu) = (x-\mu)^2/2-c, \text{ and}$$

$$\phi(x-\mu) = x-\mu.$$

The solution then is found by solving

$$\Sigma\phi(x_i-\mu) = \Sigma(x_i-\mu) = 0 \text{ or}$$

$$\hat{\mu} = \Sigma x_i/n,$$

the sample mean.

Maximum likelihood estimators may or may not be robust depending on the form of $f(x-\mu)$. However, the commonly used normal case is decidedly non-robust. To illustrate this, consider a sample of four from a normal distribution together with one outlier. Suppose the values are 2, 4, 10, 3.5, 3. The ρ functions of each x_i are given in figure 1 with c arbitrarily set to zero. The maximum likelihood estimate, $\hat{\mu}$=4.5, is being influenced greatly by x_3 = 10. On the other hand,

the median, a robust estimator, has a value of 3.5. The difficulty is that we are minimizing $\Sigma\rho(x_i-\mu)$ which is quadratic in the normal case. For an outlier, x_3 say, $\rho(x_3-\mu)$ is quite large near the center of the distribution and thus the solution $\hat{\mu}$ is pulled to the right so that $\rho(x_3-\hat{\mu})$ is not so large.

Clearly this ρ function does not result in a robust estimate. How can we robustify ρ? Intuitively, it appears that at some distance symmetric about each x_i we should make ρ increase less rapidly, perhaps even level off. Various functions have been suggested by researchers and two of these are sketched in figure 2 along with the maximum likelihood ρ. Huber's function increases quadratically along with the maximum likelihood ρ until $|x-\mu| = k_1$. For $|x-\mu| > k_1$ Huber's function increases linearly. Hampel's function follows Huber's until $|x-\mu| = k_2$ at which time $\rho(x-\mu)$ begins to level off, which it does at $|x-\mu| = k_3$. The values for k_1, k_2, and k_3 can be chosen by the researcher but reasonable values are $k_1 = 1.7$, $k_2 = 3.4$, and $k_3 = 8.5$ (Hogg 1979). The analytic representations of the functions are also given in Hogg (1979).

Robustness can be characterized more easily in terms of the ϕ functions (fig. 3). To be robust the ϕ function should at least be bounded.

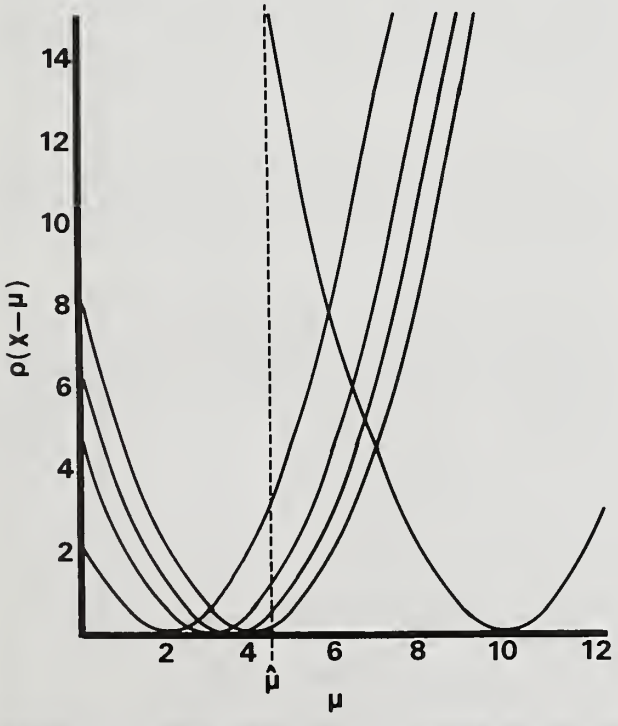


Figure 1. Normal density ρ plotted as a function of μ for each fixed x_i.

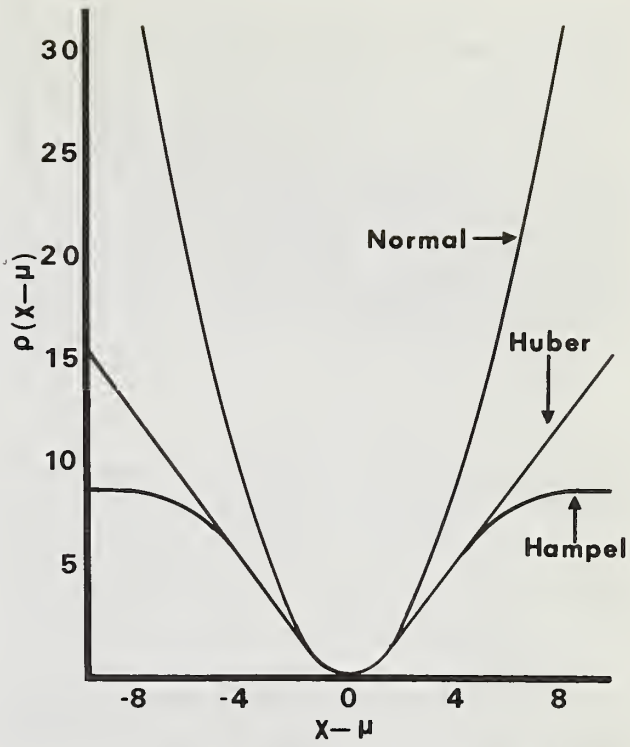


Figure 2. Normal density, Huber, and Hampel ρ's plotted as a function of $x-\mu$ with constants of $k_1 = 1.7$, $k_2 = 3.4$, and $k_3 = 8.5$.

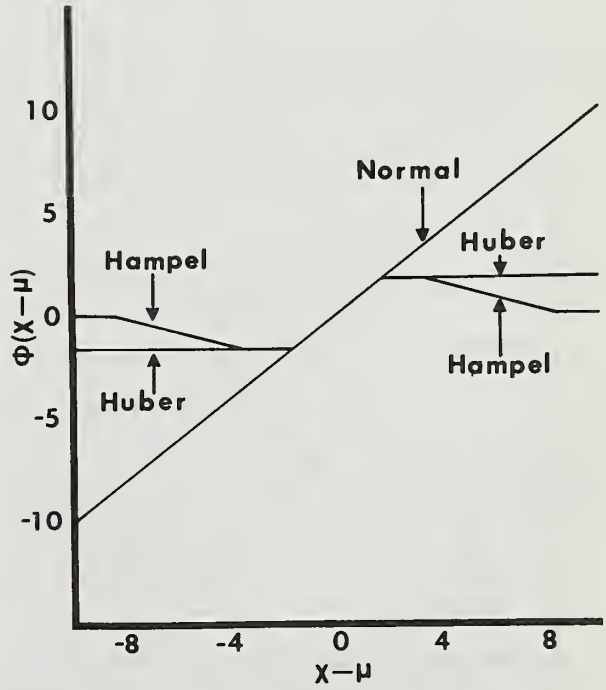


Figure 3. Normal density, Huber, and Hampel ϕ's plotted as a function of $x-\mu$ with constants of $k_1 = 1.7$, $k_2 = 3.4$, and $k_3 = 8.5$.

Note that Hampel's function even redescends to zero.

Why is the shape of the ϕ function so important? First, the influence an observation has on the estimators is proportion to ϕ for M-estimators (Hampel 1974). Thus if ϕ is unbounded as is the ϕ function in the normal case, a single "bad" value can have a great influence. Secondly, M-estimators of this type can be expressed as weighted means with the weights proportional to ϕ. That is

$$\Sigma\phi(x_i-\mu) = \Sigma w_i(x_i-\mu)$$

where $w_i = \phi(x_i-\mu)/(x_i-\mu)$.

In the normal case $w_i = 1$ for all i but the Huber and Hample estimates may have $w_i < 1$.

For simplicity we have been examining the case in which $\sigma = 1$ (i.e., the scale parameter is unity). Actually we must solve

$$\Sigma\phi[(x_i-\mu)/\sigma] = 0$$

by obtaining an ancillary estimate of σ. Generally the equation is solved iteratively starting from initial estimates, $\hat{\mu}_o$ and $\hat{\sigma}_o$. At stage j the estimate of scale is updated and the new estimate of location is then computed. The estimate of scale should be robust, such as the median absolute deviation (MAD) estimator (Mosteller and Tukey 1977).

The procedure described above can easily be extended to the regression case (e.g., Hogg 1979). Consider the model

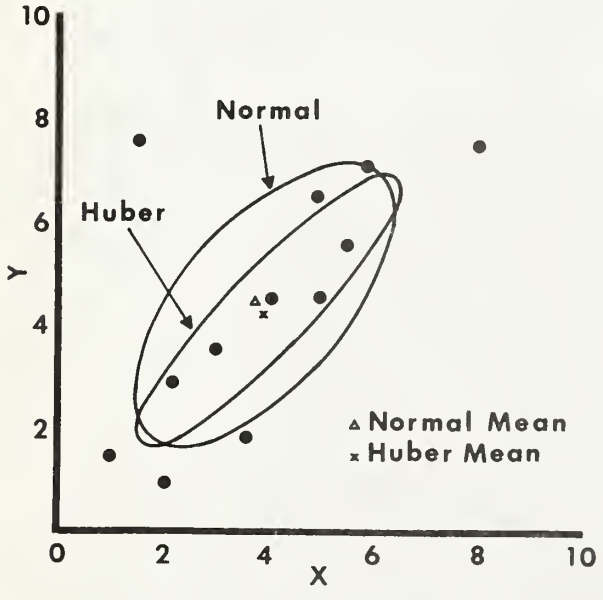


Figure 4. Example illustrating the effect of a single outlier on the mean vector and 50 percent concentration ellipse.

$$y_i = \underline{x}_i\underline{\beta}+e_i,$$

where y_i is the ith "dependent" observation, $\underline{x}_i$ is the ith 1 x p known "independent" vector, $\underline{\beta}$ is a p x 1 vector of unknown parameters, and e_i is the ith random error. We now want to solve for $\underline{\beta}$ in the p equations

$$\Sigma x_{ij}\phi[(y_i-\underline{x}_i\underline{\beta})/\sigma] = 0,\ j = 1, 2, \ldots, p,$$

where x_{ij} is the jth "independent variable" on the ith observation. Of course ancillary estimates of σ are required. The system is solved iteratively as before.

Multivariate Problem

The multivariate estimation problem is much more complicated than the univariate one. In particular, outliers in high demensional space are difficult to detect. Outliers may be apparent only on new variables which are linear combinations of the original variables. Although difficult to detect, their effect may be large. Figure 4 illustrates the effect of a single outlier on the concentration ellipse and on the sample mean vector. Notice that the sample mean vector is pulled toward the outlier. More importantly the shape of the concentration ellipse, which is determined from the sample covariance matrix, is distorted. This is important since most multivariate analyses are based on estimates of covariance matrices. The robust correlation of 0.904 more adequately expresses the relation between x and y than does the standard Pearson product-moment correlation of 0.634.

Clearly robust estimators are needed which minimize the effect of outliers. This is particularly true in the multivariate case, since variances and covariances are extremely sensitive to outliers.

Maronna (1976) and Huber (1977a, 1977b) have developed most of the theory for robustly estimating multivariate location and scale using M-estimators. Maronna characterizes M-estimators as being the solutions to the following matrix equations:

$$n^{-1}\sum_{i=1}^{n} U_1[\sqrt{d_i}](\underline{x}_i-\underline{\mu}) = 0$$

$$n^{-1}\sum_{i=1}^{n} U_2[d_i](\underline{x}_i-\underline{\mu})(\underline{x}_i-\underline{\mu})' = \underline{\Sigma}$$

where $\underline{\mu}$ is the p x 1 mean vector, $\underline{\Sigma}$ is the pxp covariance matrix, $d_i = [(\underline{x}_i-\underline{\mu})'\underline{\Sigma}^{-1}(\underline{x}_i-\underline{\mu})]^{1/2}$ is the Mahalanobis distance, and U_1 and U_2 are weight

functions. In the multivariate normal case $U_1 = U_2 = 1$ for all i and the ordinary maximum likelihood estimators result, that is $\hat{\underline{\mu}} = \Sigma \underline{x}_i/n$, and $\hat{\underline{\Sigma}} = [\Sigma(\underline{x}_i-\hat{\underline{\mu}})(\underline{x}_i-\hat{\underline{\mu}})']/n$. Huber's development of M-estimators is slightly more general than Maronna's.

We used a modified form as suggested by Randles et al. (1978). Start with $\bar{\underline{x}}$ and $\underline{S}$, the ordinary estimates. Compute

$$d_i = [(\underline{x}_i-\bar{\underline{x}})'\underline{S}^{-1}(\underline{x}_i-\bar{\underline{x}})]^{1/2}, \quad i = 1, 2, \ldots, n.$$

Replace $\bar{\underline{x}}$ and $\underline{S}$ by weighted estimates

$$\bar{\underline{x}}^* = (\Sigma w_i \underline{x}_i)/\Sigma w_i \text{ and}$$

$$\underline{S}^* = (\Sigma w_i^2(\underline{x}_i-\bar{\underline{x}}^*)(\underline{x}_i-\bar{\underline{x}}^*)')/\Sigma w_i^2$$

where $w_i = k/d_i$ if $d_i > k$ and $w_i = 1$ if $d_i \leq k$. The procedure is iterated until the change in $\bar{\underline{x}}^*$ and $\underline{S}^*$ is sufficiently small or until a predetermined number of iterations is reached.

A graph of the weight function is given in figure 5. Note that the weight is 1 for an observation unless the Mahalanobis distance is greater than k. The weight then decreases inversely with the Mahalanobis distance. If all observations are sufficiently well-behaved, then all the weights remain 1 and the ordinary estimates result.

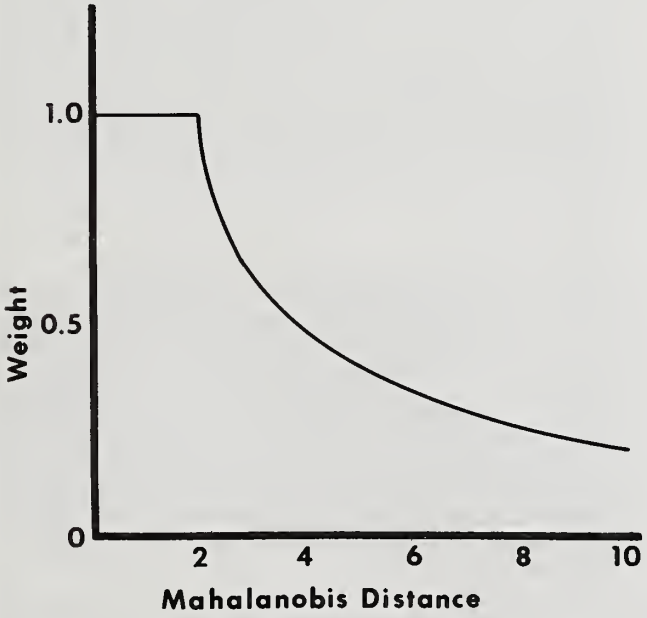


Figure 5. Weight function for estimating the mean vector and covariance matrix by Huber's method with a constant of k = 1.97.

This approach of robust estimation can be fruitfully applied in principal component analysis, factor analysis, regression, discriminant analysis, etc. The application in this paper will be in principal components and discriminant analysis.

The robust discriminant problem is discussed in Randles et al. (1978). They discuss several procedures for the two-group case. A method that is suggested naturally is to estimate the mean vector and covariance matrix separately and robustly (as above) for each group. These are then substituted into the linear or quadratic discriminant function.

Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ and $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_m$ be two independent random samples from p-variate populations. Let $\underline{z}$ be an observation from one or the other population. The Huberized linear discriminant function (HLDF) then is

$$\underline{D}_L^*(\underline{z}) = [\underline{z} - 1/2(\bar{\underline{x}}^* + \bar{\underline{y}}^*)]' \underline{S}_p^{*-1}[\bar{\underline{x}}^* - \bar{\underline{y}}^*]$$

where $\underline{x}^*$ and $\underline{y}^*$ are the robust sample mean vectors and $\underline{S}_p^*$ is the robust pooled sample covariance matrix. It is defined by

$$\underline{S}_p^* = [(n-1)\underline{S}_x^* + (m-1)\underline{S}_y^*]/(n+m-2)$$

where $\underline{S}_x^*$ and $\underline{S}_y^*$ are the robust sample covariance matrices. The robust sample discriminant coefficients are given by

$$\hat{\underline{\beta}}^* = \underline{S}_p^{*-1}(\bar{\underline{x}}^* - \bar{\underline{y}}^*).$$

Fisher's linear discriminant function (LDF) is defined similarly except that the ordinary estimates are placed in the above HLDF. Huberized or Fisherian quadratic discriminant functions can be defined easily as in Randles et al. (1978).

The geometry of robust discriminant analysis is similar to that of ordinary discriminant analysis. Concentration ellipsoids are determined by the equations of the form

$$(\underline{x}-\bar{\underline{x}}^*)'\underline{S}_p^{*-1}(\underline{x}-\bar{\underline{x}}^*) = c \text{ and}$$

$$(\underline{y}-\bar{\underline{y}}^*)'\underline{S}_p^{*-1}(\underline{y}-\bar{\underline{y}}^*) = c$$

where c is a positive constant. Probability statements can be made concerning the ellipsoids by using a chi-square distribution with p degrees of freedom as an approximation.

Stability of the Estimators

In order to assess the discriminant model it is important to assess the stability of the sample discriminant coefficients. A technique both for obtaining new estimates and for assessing their stability is the Jackknife (Mosteller and Tukey 1977). An initial step of the Jackknife procedure

called "leave-one-out" is also useful for validating the model. These procedures are not applied to the principal component coefficients in this paper although it would be possible to jackknife them.

The robust discriminant coefficients are given by $\underline{\hat{\beta}}^{*}{}' = (\hat{\beta}^{*}_{1}, \hat{\beta}^{*}_{2}, \ldots, \hat{\beta}^{*}_{p})$. The method starts by recomputing the coefficients each time after leaving out a single observation. For simplicity, let j range from 1 to N=n+m whereas i=1,2,...,p denotes the variables, i.e. $\hat{\beta}^{*}_{i(-j)}$ is the coefficient for the ith variable with the jth observation removed. $\underline{\hat{\beta}}^{*}_{(-j)}$ is used to predict the jth observation in the leave-one-out validation. Note that a prediction is being made based on values which were not used in building the model.

The jth pseudo-value for variable i is given by

$$\tilde{\beta}^{*}_{i(-j)} = N\hat{\beta}^{*}_{i} - (N-1)\hat{\beta}^{*}_{i(-j)} .$$

Then the jackknife estimator is given by

$$\tilde{\beta}^{*}_{i} = \sum_{j=1}^{N} \tilde{\beta}^{*}_{i(-j)}/N .$$

A standard deviation estimate for both $\hat{\beta}^{*}_{i}$ and $\tilde{\beta}^{*}_{i}$ is given by

$$s_{\tilde{\beta}^{*}_{i}} = \{ \sum_{j=1}^{N} (\tilde{\beta}^{*}_{i(-j)} - \tilde{\beta}^{*}_{i})^{2}/[N(N-1)] \}^{1/2} .$$

We can then test for significance since $\hat{\beta}^{*}_{i}/s_{\tilde{\beta}^{*}_{i}}$ or $\tilde{\beta}^{*}_{i}/s_{\tilde{\beta}^{*}_{i}}$ has an approximate t distribution with N-1 degrees of freedom under the null hypothesis that the coefficient is zero (Tukey 1958).

SPARROW EXAMPLE

Since 1976 data on sparrow habitats have been collected on "reclaimed" strip mines in northern West Virginia. Our purpose is to illustrate the robust procedures discussed in the previous section using a small portion of these data. In particular, habitat data of savannah sparrows (Passerculus sandwichensis) and grasshopper sparrows (Ammodramus savannarum) areused based on field work done in 1978 on the Great Mine (47.5 ha) in Preston County, West Virginia.

Ten vegetation variables were measured based on the territories of the sparrows. However, after preliminary screening, only four variables are used in this paper. They are bare ground cover (BGC), litter depth (LD), vertical diversity of grass (VH), and total grass density (TD). Methods of measurement are described by Whitmore (1979). Units of measurement are presented only in table 1. (Tables follow literature cited.) Twenty savannah sparrows and 51 grasshopper sparrows were found on this mine in 1978. However, in order to make the sample sizes equal (for convenience) only the first 20 grasshopper sparrows were used. No attempt was made to screen the data to pick the "best" observations to illustrate our techniques.

Principal Component Analyses

The effects of outliers on principal component analyses depend on their position relative to the "ellipsoidal" swarm of points in p-space. The general effect is to shift and stretch the ellipsoid in the direction of the outliers. If the outliers are along the major axes, then the ellipsoid is elongated more than it should be. On the other hand, if the outliers are orthogonal to the major axes, the ellipsoid is contracted more than it should be.

The analyses that follow estimate the mean vectors and covariance matrices by ordinary maximum likelihood (ML) adjusted to be unbiased and by Huber's method at cutoffs of k = 1.97 (H197) and k = 1.56 (H156). The cutoff at 1.97 corresponds to weighting observations less which fall in the outer 10 percent of the distribution whereas the 1.56 cutoff corresponds to those observations in the outer 30 percent of the distribution. For this data set the 1.97 cutoff should be adequate to protect against the bad effects of outliers.

The raw data are presented in table 1 and weights based on the Huber technique in table 2. The savannah sparrow data is typical of most data sets in having a few observations (about 10%) somewhat distant from the bulk, particularly observations 16 and 17. Three values are relatively bad for the grasshopper sparrow data, observations 1, 3, and 13 (table 2).

Savannah Sparrow

Summary statistics for habitat variables measured on savannah sparrows are given in table 3. The robust mean estimates were rather similar to the ML estimates although some shift is present. The robust estimate for BGC increased relative to the ML estimate, whereas the robust mean estimates for LD, VH and TD decreased (table 3). The standard deviations estimated by H197 decreased relative to those estimated by ML, except for VH. The H156 standard deviations are consistently larger than the H197 standard deviations.

The most pronounced effect caused by outlying values was on the relationships among the variables. The H197 correlation estimates were consistently less (in absolute value) than the ML estimates. The H156 estimates were not as

consistent but note that the correlation between BGC and VH is essentially zero (table 3).

In order to understand better the nature of the relationships, principal components were computed on the correlation matrix. The largest eigenvalue decreased as robustificaion increased, whereas the second largest eigenvalue increased (table 4). This indicates that the outlying observations, principally observations 1, 12, 16, and 17 (table 2), approximately lie along the first principal axis which causes the ellipsoid to be elongated with certain relationships accentuated and others obscured.

In all three techniques the percent of variation explained by the first two components was about 90% (table 4). The change of emphasis, however, is best seen by examining the eigenvectors (principal component coefficients). The first component contrasts BGC to the other variables in all cases. However, BGC becomes increasingly less important based on the robust estimates. For the ML technique, the second component seems to suggest a contrast between LD and TD. On the other hand, the robust second component suggests a linear combination of BGC and VH is important.

Grasshopper Sparrow

Summary statistics for the four variables measured on grasshopper sparrow habitat are given in table 5. Although the mean vector estimates for grasshopper sparrows were somewhat different than those for savannah sparrows, a nearly identical change occurred when the estimates were robustified. The same was true of sample standard deviations although not always with the same variables (tables 3 and 5).

For the most part, the grasshopper sparrow ML correlations were lower in absolute value than the corresponding Savannah sparrow correlations. The H197 estimates consistently indicated less strong relationships, particularly between BGC and VH. In all cases, the correlation between LD and TD remained about the same.

Results of the principal component analyses for grasshopper sparrows were analogous to those for savannah sparrows. The first two components explained about 85% of the variation with increasing importance going to the second component for the robust estimates (table 6). Again, notice the decreasing importance of BGC in the first component. In the second component, LD becomes less important whereas VH becomes more important. As with the savannah sparrow data, the robust second component is primarily a linear combination of BGC and VH (table 6).

Discriminant Analysis

The next phase of the study was to examine whether or not it is possible to discriminate between savannah and grasshopper sparrow habitat based on these four variables. Tests to determine equality of covariance matrices were not done, since it is not clear how to do this for the robust estimates. However, upon inspecting the covariance matrices, the differences between them for each estimation technique did not appear to be large enough to cause concern.

Discriminant coefficients, their standard errors, and the t-values for all methods are given in table 7. None of the coefficients were significant for any of the methods. However, in all cases the variables in the LDF model were closer to significance than the corresponding variables in the HLDF (1.97) or HLDF (1.56) models. What is the reason for this? For both species, the outliers lie along the principal axis which tends to elongate the ellipsoids, perhaps resulting in an optimistic appraisal of the LDF model's ability to discriminate. However, this better discrimination can be deceiving since estimates of the coefficients are based on outliers. Outliers in another sample might appear in totally different regions resulting in a very different discriminant model.

Predictions for the LDF, HLDF (1.97) and HLDF (1.56) models are given in table 8. A value greater or equal to zero predicts the observation to be associated with a savannah sparrow whereas a value less than zero predicts the observation to be of the grasshopper sparrow type. For the LDF model 75% of the observations are classified correctly. The corresponding result for the HLDF (1.97) model is 77.5%, and for the HLDF (1.56) model it is 65% (table 8). The LDF and HLDF (1.97) models have rather similar performances with HLDF (1.56) giving poorer results. More accurate estimates of misclassification percentages, however, were obtained from the leave-one-out predictions (table 9). These give 70, 70, and 60 correct classification percentages for LDF, HLDF (1.97), and HLDF (1.56), respectively.

We would not recommend the use of the LDF model, however, due to its dependence on outliers; the HLDF (1.97) model does as well and is more stable. As mentioned before, these data are reasonably typical with respect to outlier occurrence. Thus, a cutoff at 1.56 probably targets too many observations as outliers. This means that the HLDF (1.56) model for this data is too harsh on judging an observation to be an outlier and should not be used.

The lack of significance for the terms in the models together with the reasonably good classification indicate that models with fewer variables might discriminate nearly as well. In fact, a model based on only VH does nearly as well, but does not illustrate robustness in a multivariate setting.

Realizing that covariance matrices for each species may have been too different to justify linear discriminant analyses, ordinary and robust

quadratic discriminant analyses were done. These models did not improve the probability of correctly classifying an observation and thus are not included.

REMARKS

Outliers in multivariate data are difficult to detect if the dimensionality of the space is greater than two. It is rare that practitioners of multivariate methods attempt to adjust their analyses to minimize these effects. Partly this is due to both the unavailability of techniques to deal with multivariate outliers and the lack of computer programs. Nonetheless, outliers can have pronounced effects on the conclusions.

We suggest that both classical and robust analyses be performed on the data. If they give similar results, then use the results from the classical analysis and merely note that the robust analysis was done. If the analyses give different interpretations, it is important to understand what observations caused the discrepancies. The robust analysis then is generally to be preferred.

The M-estimation technique presented in this paper is a natural extension of maximum likelihood estimation. In fact, it is just a weighted version in which the weights are determined iteratively from the Mahalanobis distances. Thus, these robust estimators are easy to understand and interpret.

Lack of computer programs will make it difficult to perform these analyses at the present time. However, the senior author is currently developing a multivariate data analysis package. Some of these procedures should be published in the SUGI (SAS User's Group International) proceedings in 1981.

ACKNOWLEDGEMENTS

We wish to thank Gerald R. Hobbs, Harry V. Wiant, and Jake C. Rice for reviewing early drafts of this manuscript. This paper was published with the approval of the Director of the West Virginia University Agricultural and Forestry Experiment Station as Scientific Paper No. 1665.

LITERATURE CITED

Bickel, P.J. 1973. On some analogues to linear combinations of order statistics in the linear model. The Annals of Statistics 1:597-616.

Hampel, F.R. 1974. The influence curve and its role in robust estimation. Journal of the American Statistical Association 69:383-393.

Hettmansperger, T.P., and J.W. McKean. 1977. A robust alternative based on ranks to least squares in analyzing linear models. Techometrics 19:275-284.

Hogg, R.V. 1979. Statistical robustness: one view of its use in applications today. The American Statistician 33:108-115.

Huber, P.J. 1964. Robust estimation of a location parameter. The Annals of Mathematical Statistics 35:73-101.

Huber, P.J. 1977a. Robust statistical procedures. CBMS-NSF Regional Conference Series in Applied Mathematics. 56p. J.W. Arrowsmith Ltd., Bristol, England.

Huber, P.J. 1977b. Robust covariances. p. 165-191. In Gupta, S.S., and D.S. Moore, editors. Statistical decision theory and related topics II. Academic Press, New York, N.Y.

Maronna, R.A. 1976. Robust M-estimators of multivariate location and scatter. The Annals of Statistics 4:51-67.

Mosteller, F., and J.W. Tukey. 1977. Data analysis and regression. 588p. Addison-Wesley, Reading, Mass.

Randles, R.H., et al. 1978. Generalized linear and quadratic discriminant functions using robust estimates. Journal of the American Statistical Association 73:564-568.

Tukey, J.W. 1958. Bias and confidence in not-quite large samples. Abstract in the Annals of Mathematical Statistics 29:614.

Whitmore, R.C. 1979. Temporal variation in the selected habitats of a guild of grassland sparrows. Wilson Bulletin 91:592-598.

Table 1. Measurements of four habitat variables for savannah and grasshopper sparrows.[1]

	Savannah sparrow				Grasshopper sparrow			
Obs	BGC	LD	VH	TD	BGC	LD	VH	TD
1	24.3	1.02	0.98	897	19.8	6.95	1.21	899
2	36.4	1.45	0.74	449	38.7	1.20	0.57	174
3	19.1	1.02	1.00	579	4.5	3.55	0.87	364
4	42.5	0.26	0.89	197	55.0	0.47	0.54	133
5	58.4	0.18	0.56	113	33.0	2.87	0.60	418
6	15.3	1.97	1.04	432	0.5	2.95	0.73	714
7	48.8	0.27	0.52	145	33.0	0.51	0.71	222
8	50.0	0.47	0.64	212	20.5	2.41	0.79	743
9	18.0	3.31	0.83	830	22.3	0.34	0.81	188
10	44.4	2.67	0.88	381	10.5	2.00	0.83	454
11	20.0	3.15	1.14	826	24.9	1.18	0.77	404
12	14.8	1.18	0.77	859	55.5	0.53	0.61	95
13	13.4	3.10	0.92	604	66.1	0.56	0.00	154
14	40.4	0.53	0.44	210	22.9	0.48	0.32	274
15	10.6	2.60	1.00	480	20.5	1.66	0.51	308
16	0.7	6.30	1.30	1208	24.3	1.20	0.45	271
17	11.3	4.46	0.92	468	14.2	1.76	1.21	629
18	68.5	0.87	0.84	120	22.0	2.77	0.57	431
19	1.4	3.55	1.31	1090	58.5	0.41	0.56	127
20	13.7	2.42	1.07	810	14.2	2.31	0.50	451

[1]Variables are BGC - bare ground cover (%); LD - litter depth (cm); VH - vertical diversity of grass (Shannon-Weaver index); and TD - total grass density (hits/unit transect).

Table 2. Final weights for each four-variate observation using Huber's robust procedure with cutoffs of 1.97 and 1.56.

	Final Weights			
	Savannah sparrow		Grasshopper sparrow	
Obs	H197	H156	H197	H156
1	0.786	0.537	0.319	0.271
2	1	1	1	1
3	1	0.736	0.495	0.335
4	0.960	0.640	1	0.987
5	1	0.986	1	0.887
6	0.924	0.648	0.871	0.608
7	1	1	1	1
8	1	1	0.650	0.449
9	0.964	0.639	0.814	0.604
10	1	0.900	1	0.984
11	1	1	1	1
12	0.757	0.540	1	0.888
13	1	1	0.479	0.328
14	1	1	0.926	0.726
15	1	0.765	1	1
16	0.534	0.408	1	1
17	0.679	0.471	0.651	0.487
18	0.800	0.802	1	1
19	1	0.780	1	0.906
20	1	1	1	1

Table 3. Sample mean vectors, standard deviations, and correlations for four variables measured on savannah sparrow habitat using maximum likelihood and Huber's robust procedure with cutoffs of 1.97 and 1.56.

	Sample mean vectors			
	BGC	LD	VH	TD
ML	27.60	2.039	0.890	545.5
Huber (1.97)	28.13	1.940	0.880	530.6
Huber (1.56)	29.63	1.880	0.866	510.6
	Sample standard deviations			
ML	19.50	1.625	0.235	334.4
Huber (1.97)	19.16	1.325	0.248	301.2
Huber (1.56)	23.14	1.329	0.328	309.7

	Sample correlations					
	BGC vs LD	BGC vs VH	BGC vs TD	LD vs VH	LD vs TD	VH vs TD
ML	-0.728	-0.724	-0.831	0.715	0.699	0.761
Huber (1.97)	-0.715	-0.500	-0.757	0.608	0.698	0.723
Huber (1.56)	-0.421	0.032	-0.441	0.693	0.780	0.767

Table 4. Sample eigenvalues and eigenvectors based on correlation matrices for four variables measured on savannah sparrow habitat using maximum likelihood and Huber's robust procedure with cutoffs of 1.97 and 1.56.

Sample eigenvalues

	Component			
Estimator	1	2	3	4
ML	3.230	0.326	0.283	0.160
(Cum. Percent)	80.8	88.9	96.0	100.0
H197	3.007	0.516	0.311	0.166
(Cum. Percent)	75.2	88.1	95.8	100.0
H156	2.641	1.028	0.234	0.097
(Cum. Percent)	66.0	91.7	97.6	100.0

Sample eigenvectors

Estimator	Component	BGC	LD	VH	TD
ML	1	0.509	-0.485	-0.495	-0.510
	2	-0.360	-0.791	-0.096	0.485
	3	-0.425	0.336	-0.838	0.068
	4	-0.656	-0.160	0.211	-0.707
H197	1	0.496	-0.503	-0.468	-0.531
	2	-0.578	0.232	-0.779	-0.073
	3	-0.350	-0.812	-0.026	0.466
	4	-0.546	-0.183	0.416	-0.704
H156	1	0.292	-0.565	-0.505	-0.583
	2	-0.857	0.022	-0.515	-0.004
	3	-0.150	-0.809	0.210	0.528
	4	-0.398	-0.158	0.659	-0.618

Table 5. Sample mean vectors, standard deviations, and correlations for four variables measured on grasshopper sparrow habitat using maximum likelihood and Huber's robust procedure with cutoffs of 1.97 and 1.56.

Sample mean vectors

	BGC	LD	VH	TD
ML	28.04	1.806	0.658	372.6
Huber (1.97)	28.63	1.591	0.635	345.7
Huber (1.56)	28.99	1.569	0.630	337.6

Sample standard deviations

ML	18.24	1.576	0.274	227.0
Huber (1.97)	19.32	0.992	0.225	176.2
Huber (1.56)	20.84	1.048	0.276	182.7

Sample correlations

	BGC vs LD	BGC vs VH	BGC vs TD	LD vs VH	LD vs TD	VH vs TD
ML	-0.518	-0.542	-0.673	0.563	0.811	0.640
Huber (1.97)	-0.473	0.039	-0.537	0.266	0.779	0.478
Huber (1.56)	-0.171	0.381	-0.194	0.438	0.836	0.622

Table 6. Sample eigenvalues and eigenvectors based on the correlation matrices for four variables measured on grasshopper sparrow habitat using maximum likelihood and Huber's robust procedure with cutoffs of 1.97 and 1.56.

Sample eigenvalues

	Component			
Estimator	1	2	3	4
ML	2.883	0.502	0.457	0.158
(Cum. percent)	72.1	84.6	96.0	100.0
H197	2.349	1.048	0.442	0.161
(Cum. percent)	58.7	84.9	96.0	100.0
H156	2.279	1.290	0.332	0.099
(Cum. percent)	57.0	89.2	97.5	100.0

Sample eigenvalues

Estimator	Component	BGC	LD	VH	TD
ML	1	0.471	-0.506	-0.473	-0.547
	2	0.704	0.631	-0.235	0.225
	3	-0.477	0.151	-0.847	0.182
	4	-0.236	0.568	0.065	-0.786
H197	1	0.437	-0.572	-0.318	-0.617
	2	0.587	-0.082	0.801	0.078
	3	-0.641	-0.645	0.410	-0.067
	4	-0.231	0.500	0.297	-0.780
H156	1	0.025	-0.589	-0.502	-0.633
	2	0.843	-0.209	0.474	-0.148
	3	-0.479	-0.632	0.602	0.092
	4	0.245	-0.458	-0.402	0.754

Table 7. Four-variate discriminant coefficient estimates and standard errors for Fisher's linear discriminant model and Huber's robust linear discriminant models.

	LDF			
	BGC	LD	VH	TD
$\hat{\beta}$	0.0656	-0.525	5.40	0.00427
$s_{\hat{\beta}}$	0.0431	0.668	2.97	0.00282
$t_{\hat{\beta}}$	1.52	-0.79	1.82	1.51
	HLDF (1.97)			
	BGC	LD	VH	TD
$\hat{\beta}^*$	0.0249	-0.336	3.45	0.00338
$s_{\hat{\beta}^*}$	0.0303	0.655	2.45	0.00340
$t_{\hat{\beta}^*}$	0.82	-0.51	1.41	0.99
	HLDF (1.56)			
	BGC	LD	VH	TD
$\hat{\beta}^*$	-0.00402	-0.683	2.38	0.00308
$s_{\hat{\beta}^*}$	0.0294	0.682	3.09	0.00369
$t_{\hat{\beta}^*}$	-0.14	-1.00	0.77	0.83

Table 8. Discriminant predictions based on Fisher's linear discriminant model and Huber's robust linear discriminant models.

	Predictions					
	Savannah sparrow			Grasshopper sparrow		
Obs	LDF	(1.97) HLDF	(1.56) HLDF	LDF	(1.97) HLDF	(1.56) HLDF
1	3.22	2.47	2.51	1.07	1.16	-0.96
2	0.59	0.28	0.22	-1.23	-1.09	-0.88
3	1.64	1.33	1.60	-2.27	-1.05	-1.04
4	1.35	0.50	0.58	-0.11	-0.68	-0.64
5	0.29	-0.50	-0.47	-1.27	-0.87	-1.17
6	0.48	0.56	0.61	-1.48	-0.25	0.13
7	-0.47	-0.80	-0.49	-0.28	-0.36	0.10
8	0.44	-0.20	-0.14	0.56	0.73	0.65
9	0.51	0.80	0.41	-0.49	-0.34	0.39
10	0.94	0.32	-0.53	-0.90	-0.22	0.17
11	2.39	1.96	1.24	-0.06	0.04	0.38
12	1.22	1.33	1.82	0.11	-0.58	-0.63
13	-0.16	0.30	0.09	-2.26	-2.23	-1.97
14	-1.31	-1.15	-0.62	-2.81	-1.77	-0.61
15	-0.17	0.26	0.25	-2.41	-1.45	-0.85
16	1.96	2.26	0.72	-2.40	-1.54	-0.80
17	-1.59	-0.67	-1.25	2.27	1.85	1.77
18	2.13	0.51	-0.29	-2.05	-1.17	-1.09
19	3.00	2.84	2.26	0.23	-0.53	-0.58
20	1.91	1.75	1.54	-2.61	-1.38	-0.85
Mis-classified	5	5	7	5	4	7

Table 9. Discriminant leave-one-out predictions based on Fisher's linear discriminant and Huber's robust linear discriminant models.

	Leave-One-Out Predictions					
	Savannah sparrow			Grasshopper sparrow		
Obs	LDF	(1.97) HLDF	(1.56) HLDF	LDF	(1.97) HLDF	(1.56) HLDF
1	3.32	2.33	2.30	3.85	1.79	-0.37
2	0.55	0.24	0.17	-1.17	-1.01	-0.78
3	1.53	1.26	1.48	-2.13	-1.01	-0.82
4	1.17	0.23	0.33	0.08	-0.45	-0.36
5	0.07	-0.79	-0.78	-1.17	-0.55	-0.91
6	0.25	0.32	0.31	-1.32	0.01	0.29
7	-0.66	-1.08	-0.72	-0.16	-0.14	0.40
8	0.32	-0.35	-0.30	0.82	0.99	0.84
9	0.31	0.51	0.11	-0.14	0.15	0.97
10	0.76	-0.06	-0.89	-0.78	-0.01	0.49
11	2.33	2.05	1.31	0.01	0.15	0.55
12	0.83	0.88	1.47	0.36	-0.34	-0.39
13	-0.26	0.20	-0.06	-1.96	-2.49	-1.69
14	-1.64	-1.47	-0.86	-2.78	-1.53	-0.45
15	-0.43	0.04	-0.08	-2.36	-1.41	-0.80
16	1.68	1.64	0.10	-2.35	-1.47	-0.76
17	-2.72	-2.02	-2.42	3.46	2.41	2.33
18	1.92	0.18	-0.67	-1.98	-0.97	-0.83
19	3.01	3.13	2.29	0.52	-0.23	-0.29
20	1.84	1.86	1.69	-2.56	-1.24	-0.67
Mis-classified	5	6	9	7	6	7

DISCUSSION

STEVEN PARREN: In addition to the elimination of outliers, does robust DFA have an advantage over transformations of habitat variables as we measure (and perceive) these variables in the field?

E. JAMES HARNER: The use of robust estimates does not preclude the use of transformations but it often makes it unnecessary to transform. An outlier can make it appear that a log (say) transform is necessary. I prefer not to transform unless absolutely necessary because of the difficulty of interpretation.

STEVEN PARREN: Are homogenous covariance matrices ever found in ecological studies?

E. JAMES HARNER: Yes, within reasonable statistical variability. In most cases, however, they are probably both biologically and statistically different. We need more research on the effect of not satisfying this assumption on the linear discriminant function. I believe A.P. Dempster (1969. Elements of continuous multivariate analysis. 388 p. Addison-Wesley, Reading, Mass.) somewhat quantifies the "difference" in covariance matrices.

STEVEN PARREN: What effects do transformations of variables used in multivariate space (DFA) have on the ecological interpretation of habitat variables?

E. JAMES HARNER: If the object is strictly classification then by using, for example, the power family of transformations causes few conceptual difficulties. However, if you desire to "interpret" the coefficients I believe it does. The robust approach can sometimes make it unnecessary to transform.

LESLIE MARCUS: Standard errors of your jackknifed coefficients are large, which support the idea that it is difficult to interpret coefficients. Please comment.

E. JAMES HARNER: The t-tests based on these standard errors are not significant, but the robust standard errors are on the whole somewhat smaller than those of the standard analysis. We did not use variable selection in determining this model. In order to complete the model-building process, it would be necessary to eliminate one or more variables from the discriminant analysis to remove the redundancies.

USE OF DISCRIMINANT ANALYSIS AND OTHER STATISTICAL METHODS IN ANALYZING MICROHABITAT UTILIZATION OF DUSKY-FOOTED WOODRATS[1]

Janet I. Cavallaro[2], John W. Menke[3], and William A. Williams[4]

Abstract.--In the past 10 years, discriminant analysis has become an increasingly used statistical method in small mammal studies. Although originally developed as a classification procedure, it has been used by ecologists and wildlife biologists for a variety of other purposes. No generally accepted methods, however, have been developed for interpreting the discriminant functions used for these alternative purposes.

Using the dusky-footed woodrat (Neotoma fuscipes) as an example, we present a method for interpreting discriminant functions when the purpose is both to characterize microhabitats used by a species and to explain why individuals of a species occur in some microhabitats and not in others. Three types of discriminant functions were identified from which hypotheses of different forms were developed. More complete interpretation of two-group discriminant functions can be made if a multiple regression is calculated for the dependent variable, presence or absence of woodrats and if partial correlations are calculated for the independent variables. Since multiple regression coefficients are unstable if multicollinearity exists among the independent variables, the variance inflation factor was calculated and ridge regression was performed so that the effect of intercorrelated variables could thereby be examined and unstable variables eliminated wherever necessary.

Key words: California chaparral; multicollinearity; multiple regression; partial correlation; ridge regression; variance inflation factor; wildlife habitat analysis.

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Ph.D. Graduate Student, Department of Forestry and Resource Management, University of California, Berkeley, CA 94720.

[3]Associate Professor, Department of Agronomy and Range Science, University of California, Davis, CA 95616.

[4]Professor, Department of Agronomy and Range Science, University of California, Davis, CA 95616.

INTRODUCTION

Discriminant analysis was developed as a classification procedure for assigning unknown specimens to one of two or more groups. In small mammal studies, however, discriminant analysis has also been used for other purposes: 1) evidence of niche separation, 2) hypothesis formulation about the mechanism of coexistence of sympatric species, and 3) characterization of microhabitats.

M'Closkey (1976) and Holbrook (1978) used discriminant analysis to provide evidence of niche separation of small mammal species by equating structural variables in discriminant functions to niche dimensions that separated the niches of small mammal species. M'Closkey recommended caution in equating statistical and biological patterns of microhabitat use, but then went on to suggest that

> "the statistical patterns of separation reflect ecological and evolutionary adjustments to inter-specific competition. Structural habitat division by rodents is either the means by which co-existence is achieved or is correlated with other niche dimensions (e.g., food) responsible for coexistence."

Unless hypotheses are developed from the discriminant function to explain how species partition the habitat according to structural variables, however, little new insight is achieved.

In other work, M'Closkey and Fieldwick (1975) used discriminant analysis to develop hypotheses to explain how two sympatric species could coexist. They hypothesized that since the discriminant function contained structural habitat variables that were characteristic structural features of optimal habitats of Microtus pennsylvanicus or Peromyscus leucopus, the two species were sympatric because animals of each species could find microhabitats which corresponded to their optimal habitat. They observed that foliage height diversity and tree basal area in Peromyscus microhabitats were respectively 1.5 and 4 times greater than in Microtus microhabitats, and the depth of the grassmat in Microtus microhabitats was twice as deep as in the Peromyscus microhabitats.

Dueser and Shugart (1978) used discriminant analysis to characterize the microhabitats of small mammal species and to explain why individuals of a species occurred in some microhabitats and not in others. By restricting their analysis to significant univariate variables they were able to conclude that species X occurred where more of habitat variable A occurred, where habitat variable B was larger, and so forth. They used the simple correlation between the function and each variable in the function to measure the "relative contribution of the variable to the power of the function to discriminate between groups". Simple correlation coefficients, however, are inappropirate whenever more than one independent variable is involved since they only show which univariate variables would separate the groups similarly to the function. The relative magnitude of the correlation coefficient corresponds exactly to relative value of the univariate F-ratios (Marascuilo and Levin[5]).

Using live-trap data collected on dusky-footed woodrats in chamise-wedgeleaf ceanothus (Adenostoma fasciculatum-Ceanothus cuneatus) chaparral in northern California, we present a method for interpreting discriminant functions that includes: 1) determination of the type of discriminant function to be used, 2) assessment of multicollinearity with the variance inflation factor and ridge regression, 3) use of multiple regression and partial correlation analysis, and 4) development of hypotheses to explain why individuals of a species frequent some microhabitats and not others.

FIELD METHODS

Woodrats were trapped on 0.5-ha and 1.0-ha grids in 14- and 25-year old chamise-ceanothus chaparral, respectively, for periods of 6 nights. Fifty-two structural, botanical, and physical habitat variables describing each microhabitat were measured using point sampling and nearest neighbor methods (Cavallaro 1978).

INTERPRETING DISCRIMINANT FUNCTIONS

Discriminant functions can serve as the basis for characterizing habitats used by individuals of a species. The characterization, however, is much more useful if researchers hypothesize why a species uses habitats with particular characteristics. Partial correlation analysis can assist in hypothesis development. Hereafter, we present our approach to interpreting discriminant functions in wildlife habitat studies using the woodrat as an example.

Selection of Discriminant Function Type

We have identified three types of discriminant functions that help both to characterize the habitats used by a species and to generate hypotheses explaining that habitat use. The types are distinguished on the basis of whether or not the function contains variables which as univariate variables identify a difference between where individuals of a species do and do not occur. The three types of functions are as follows: Type I contains only significant univariate variables, Type II contains both significant and nonsignificant univariate

---

[5]Textbook in preparation, Marascuilo, L.A., and J. Levin, Department of Education, University of California, Berkeley.

Table 1. Variables included in two discriminant functions that separated trap sites used and not used by dusky-footed woodrats on the 14- and 25 year-old plots; also shown are univariate statistics and variance inflation factors.

Variable	Discriminant function type	Univariate		Variance inflation factor	
		F	R^2	DF Type I	DF Type II
14-year-old plot					
1. Other shrub species density	I, II	**	0.175	0.4	121.4
2. Other shrub species live leaf density	I	**	0.135	3.4	
3. Other shrub species live stem density	I, II	**	0.197	15.5	116.0
4. Other shrub species dead stem density	I	**	0.159	10.1	
5. Chamise dead stem density	I	**	0.136	2.8	
6. Chamise density	I, II	*	0.099	2.8	2.0
7. Total vegetation cover	II		0.004		1.7
8. Vertical canopy density 100-150 cm aboveground	II		0.021		3.0
9. Vertical canopy density 150-200 cm aboveground	II		0.009		2.0
10. Total stem (< 0.5 cm in diameter) density	II		0.022		2.3
11. Ceanothus live leaf density	II		0.007		1.4
12. Yerba santa live leaf density	II		0.026		1.1
Type I discriminant function**, R^2 = 0.29					
Type II discriminant function**, R^2 = 0.53					
25-year-old plot					
13. Other ground cover	I, II	**	0.103	1.1	1.2
14. Live stem (< 0.5 cm in diameter) density	I, II	*	0.070	1.1	1.2
15. Live stem (1.0-2.5 cm in diameter) density	I, II	*	0.064		
16. Chamise live leaf density	I, II	*	0.068	1.1	7.1
17. Vertical canopy density 0-2.5 cm aboveground	II		0.016		1.2
18. Live leaf density	II		0.026		6.7
19. Total stem (0.5-1.0 cm in diameter) density	II		0.043		1.1
Type I discriminant function**, R^2 = 0.21					
Type II discriminant function**, R^2 = 0.40					

*P < 0.05, **P < 0.01

variables, and Type III contains only nonsignificant univariate variables. Hypotheses based on Type I discriminant functions would state that where woodrats occur variables x, y, and z, respectively, are less, greater, and less than where woodrats do not occur for some proposed reason. In contrast, hypotheses based on Type III discriminant functions would state that where woodrats occur X = value D, Y = value E, and Z = value F, and this particular combination of variable values is important for some proposed reason even though similar values of x, y, and z can be found in microhabitats not used by woodrats. Hypotheses based on Type II discriminant functions would state that where woodrats occur variable X is greater, Y = value H, and Z = value I for a stated reason which must explain why more of X is necessary and why variables Y and Z have their particular values even though similar values of Y and Z can be found in microhabitats not used by woodrats.

It becomes apparent that hypotheses developed from Type II or III discriminant functions are much more difficult to justify because the functions would indicate that animals are using extremely specific microhabitats. Those discriminant functions may often have questionable ecological significance. For comparison, however, we will include Type I and II discriminant functions in our discussion.

If a satisfactory hypothesis can be developed from a Type I discriminant function, selection of variables is a simple matter. All significant univariate variables are used. Selection of variables to include in a Type II or Type III discriminant function is more difficult. A useful strategy is to use those variables that increment the coefficient of determination, R^2, by a specified minimum amount (Marascuilo and Levin[5]); we used a 2% increment for Type II and III discriminant functions.

Discriminant Functions

For the 14-year-old plot, the Type I discriminant function had a R^2 value of 29% and contained all significant univariate variables except one, other shrub species live leaf density (2) (table 1). The Type II discriminant function had a R^2 value of 53% and contained in addition to all nonsignificant univariate variables listed in table 1 three density variables, other shrub species (1), other shrub species live stems (3), and chamise (6). For the 25-year-old plot, the Type I discriminant function had a R^2 value of 21% and contained all significant univariate variables. The Type II discriminant function had a R^2 value of 40% and contained all nonsignificant univariate variables listed in table 1 as well as all significant univariate variables.

The coefficient of determination, R^2, is considered a useful statistic because a function may be significant and yet very little variation in the function may be associated with the variables in the discriminant function. In such cases, the function contributes little information. In the field of educational psychology, R^2 values less than 10% are considered too inconsequential to be reported.[6]

Knowing the importance of each variable in the discriminant functions would be extremely useful for developing hypotheses. Fortunately, the two-group case of discriminant analysis can be done as a multiple regression, with the dependent variable being presence or absence of woodrats; and thereafter, partial correlations can be calculated to assess the importance of habitat variables. Partial correlations are a measure of association between each independent variable and the dependent variable in the multiple regression.

Assessment of Multicollinearity in Multiple Regression

In data sets containing measurements of uncontrolled variables, as is usually the case in wildlife habitat studies, multicollinearity often exists among variables and can cause ambiguous results in ordinary multiple regression (Hoerl and Kennard 1970). Resulting regression coefficients can be inflated in absolute value or may even have the wrong sign. The possibility that such difficulties will occur increases as independent variables diverge from orthogonality. Thus, it is desirable to test for multicollinearity among independent variables.

Calculation of the variance inflation factor (VIF) is the preferred method for determining whether a multicollinearity problem exists, because the VIF will detect relations that exist among several variables that might not be evident from examination of simple correlations (Williams et al. 1979). VIF equals $1/(1-R_i^2)$ where R_i^2 is the multiple correlation coefficient of one independent variable with all other independent variables. As R_i^2 approaches 0 (orthogonality) VIF approaches 1; whereas, as R_i^2 approaches 1 (variables are highly intercorrelated) VIF approaches infinity. VIF values exceeding 10 are considered likely to cause problems in estimating regression coefficients (Chatterjee and Price 1977). "Tolerance" is the reciprocal of the VIF, and the equivalent critical value is 0.1.

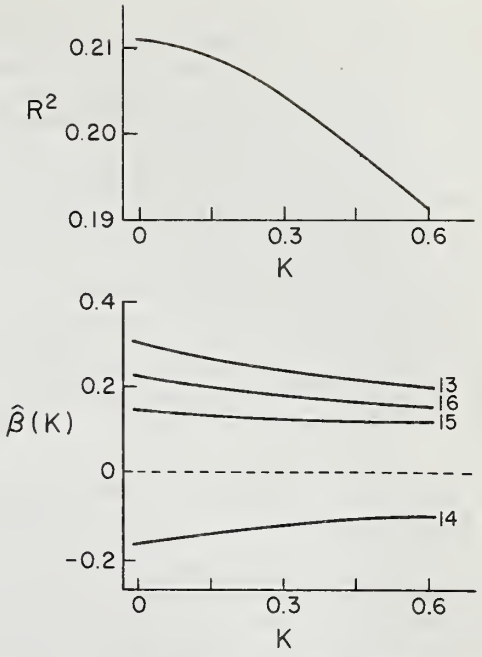


Figure 1. Ridge trace for variables in the Type I discriminant function for the 25-year-old plot. Variables as defined in table 1.

One or more variables in the two discriminant functions for the 14-year-old plot has a VIF value which exceeds 10, suggesting that for those variables the regression coefficients probably are not stable (table 1). In contrast, neither function for the 25-year-old plot contained any variables with a VIF greater than 10, thus all those variables can be used in calculating a multiple regression and partial correlations.

Ridge regression was developed to alleviate problems from intercorrelated independent variables by including a bias factor k when regression coefficients are estimated (Hoerl and Kennard 1970). The first step in ridge regression involves making a ridge trace by adding increments of k over the range of 0 to 1 and plotting the respective standardized partial regression coefficients $\hat{\beta}(k)$ against k. As k increases the coefficient of determination, R^2, decreases.

The ridge trace for variables contained in

[6] Personal communication with L.A. Marascuilo, Professor of Statistics, Department of Education, University of California, Berkeley.

the Type I discriminant functions for the 25 year-old plot (fig. 1) shows that the regression coefficients are stable as would be expected since no variables in that discriminant function had a VIF value greater than 10. As k increases, the standardized regression coefficients $\hat{\beta}(k)$ maintain both their absolute value and their values relative to each other quite well. In contrast, in the ridge trace for variables contained in the Type I discriminant function for the 14 year-old plot (fig. 2) variables 2 and 4, other shrub species live leaf density and other shrub species dead stem density, were negative for k = 0 and became positive for k > 0.13, indicating that those variables are highly unstable. Variable 3, other shrub species live stem density, was high and positive at k = 0, but decreased rapidly as k increased, suggesting that variable 3 is also unstable.

The ridge trace shows either what k value to use, if a multiple regression is desired for a particular combination of variables (a value which stabilizes the regression coefficients while not greatly decreasing the R^2 value), or which variables to eliminate because they are unstable. The ridge trace shows visually what the VIF values suggest.

Variables lacking stability were eliminated from the Type I discriminant function for the 14-year-old plot leaving other shrub species live stem density (3) and chamise dead stem density (5). Both had a VIF of 1.3 and stable regression coefficients upon calculation (fig. 3).

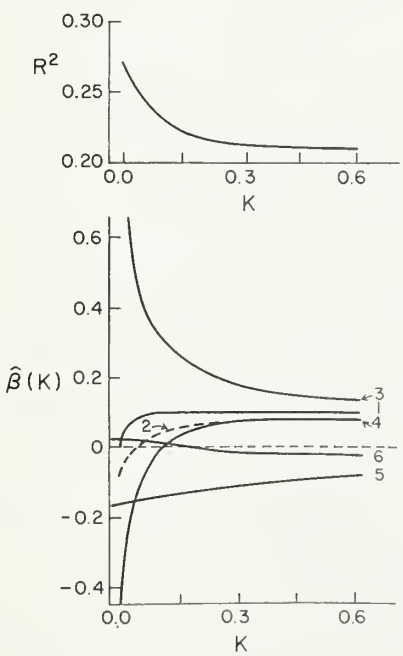


Figure 2. Ridge trace for variables in the Type I discriminant function for the 14-year-old plot. Variables are defined in table 1.

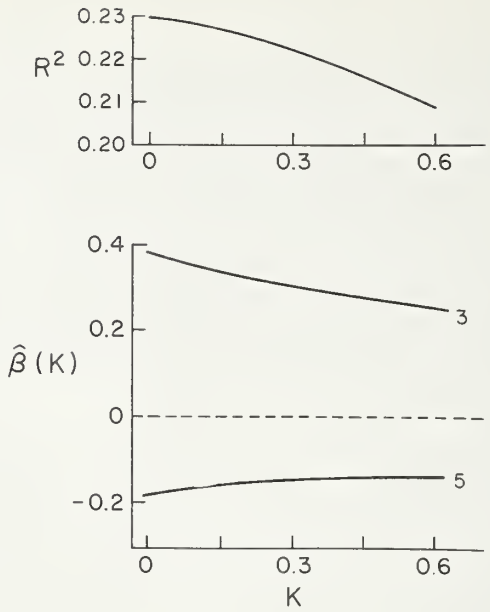


Figure 3. Ridge trace for variables originally in the Type I discriminant function for the 14-year-old plot which remained after assessing intercorrelation among variables. Variables are defined in table 1.

The ridge trace for variables in the Type II discriminant function for the 14-year-old plot is shown in figure 4. Variable 1, other shrub species density, was strongly negative for k = 0 but became positive for k > 0.06. In contrast, variable 3, other shrub species live stem density was strongly positive for k = 0 but decreased rapidly in value for k > 0. For k = 0 these two variables had opposite signs, even though they had a simple correlation of 0.99 and therefore represent the same underlying variable. These results demonstrate how multicollinear variables can greatly affect the stability of regression coefficents. After eliminating two variables in the Type II discriminant function, the VIF for all variables was less than 3.0, and a new ridge trace of the restricted variables show that the variables were much more stable (fig. 5).

Microhabitat Use Hypotheses

Based on the Type I discriminant function for the 14 year-old plot, microhabitats used by woodrats can be characterized generally as having greater density of other shrub species and lesser density of chamise, and specifically as having greater density of live leaves, live stems, and dead stems of other shrub species and less density of dead chamise stems, than microhabitats not used. This combination of variables was associated with 29% of variability in woodrat occurrence (table 1 and 2). The reason woodrats

occur in these microhabitats can be hypothesized from partial correlation coefficients. Although a correlation does not prove cause and effect, it does suggest variables to consider in trying to establish a cause and effect relationship. A positive relationship exists between woodrat presence and amount of live stems of other shrub species which included deerbrush (Ceanothus integerrimus), scrub oak (Quercus dumosa), poison oak (Toxicodendron radicans), and toyon (Heteromeles arbutifolia) with scrub oak being by far most abundant. In contrast, no relationship existed between woodrat presence and amount of dead chamise stems. We, therefore, suggest that woodrats occurred in particular microhabitats because of presence of other shrub species, rather than because of lesser amounts of chamise. This hypothesis agrees with the observation that oak species are the most important variable in determining presence of dusky-footed woodrats (Linsdale and Tevis 1951).

Based on the Type II discriminant function for the 14-year-old plot, microhabitats used by woodrats can be characterized generally as having greater density of other shrub species and less density of chamise, and specifically as having greater density of live stems of other shrub species, total vegetation cover of 80.9%, vertical canopy density of 0.6 and 0.0, respectively, for layers between 100-150 cm and between 150-200 cm above the ground, total density of 1.5 for stems less than 0.5 cm in diameter, and density of live leaves of 0.1 and 0.0 respectively for ceanothus and yerba santa. In contrast, microhabitats where woodrats did not occur can be characterized as having less density of other shrub species, particularly live stems of other shrub species, greater density of chamise, total vegetation cover of 78.7%, vertical canopy density of 0.5 and 0.1, respectively, for layers between 100-150 cm and between 150-200 cm above the ground, total density of 1.7 for stems less than 0.5 cm in diameter, and density of live leaves of 0.1 and 0.0, respectively for ceanothus and yerba santa. We do not think this function provides a good characterization of microhabitats used and not used by woodrats because we can give no reason why woodrats would select such specific microhabitats. Why, for example, should they distinguish between 80.9 and 78.7% total vegetation cover, and so forth? Consequently, we suspect that the Type II discriminant function has statistical significance in this case but lacks ecological validity.

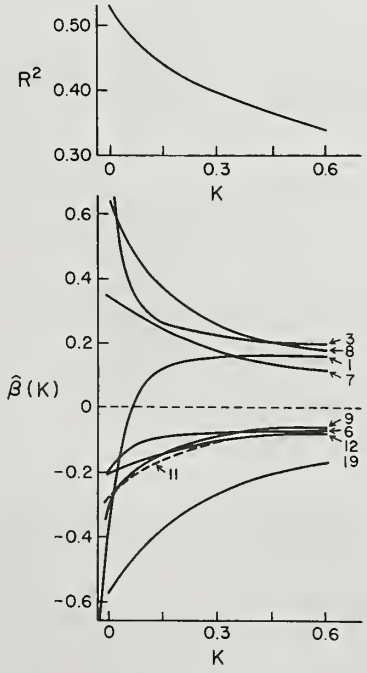


Figure 4. Ridge trace for variables in the Type II discriminant function for the 14-year-old plot. Variables are defined in table 1.

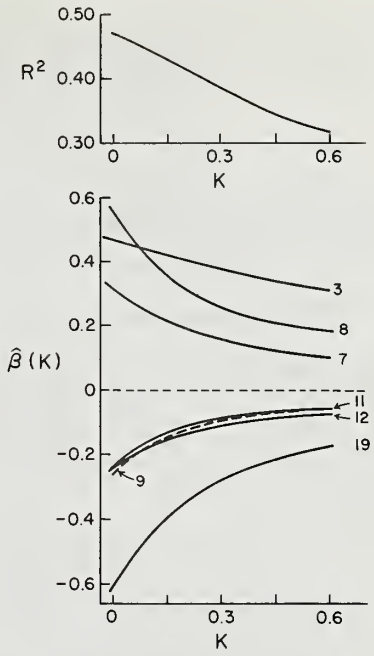


Figure 5. Ridge trace for variables originally in the Type II discriminant function for the 14-year-old plot which remained after assessing the intercorrelation among the variables. Variables are defined in table 1.

On the 25-year-old plot, where other shrub species were essentially absent (they accounted for only 1.5% of the shrub cover in contrast to 6% on the 14-year-old plot) and where all other species shrub cover was produced by deerbrush, woodrats did not occur in microhabitats with greater density of other shrub species. On the 25-year-old plot, they occurred instead where more other ground cover was present, density of live chamise leaves was greater, density of live stems between 1.0 and 2.5 cm in diameter was greater, and density of live stems less than 0.5 cm in

Table 2. Mean values (percent) for all variables in Type I and II discriminant functions and partial correlations for variables in Type I and II multiple regressions for the 14- and 25-year-old plots.

Variable	Microhabitat Used	Microhabitat Not Used	Partial Correlation
----------14-year-old plot----------			
Type I discriminant function			
1. Other shrub species density	0.6	0.0	
2. Other shrub species live leaf density	0.1	0.0	
3. Other shrub species live stem density	0.2	0.0	0.33**
4. Other shrub species dead stem density	0.3	0.0	
5. Chamise dead stem density	0.6	0.9	-0.20
6. Chamise density	1.4	2.1	
Type II discriminant function			
1. Other shrub species denisty	0.6	0.0	
3. Other shrub species live stem density	0.2	0.0	0.48**
6. Chamise density	1.4	2.1	
7. Total vegetation cover	80.9	78.7	0.30*
8. Vertical canopy density 100-150 cm aboveground	0.6	0.5	0.41**
9. Vertical canopy density 150-200 cm aboveground	0.0	0.1	-0.28**
10. Total stem (<0.5 cm in diameter) density	1.5	1.7	-0.53**
11. Ceanothus live leaf density	0.1	0.1	-0.30*
12. Yerba santa live leaf density	0.0	0.0	-0.33*
----------25-year-old plot----------			
Type I discriminant function			
13. Other ground cover	2.1	0.3	0.28**
14. Live stem (<0.5 cm diameter) density	0.2	0.5	-0.20
15. Live stem (1.0-2.5 cm in diameter) density	0.1	0.0	0.11
16. Chamise live leaf density	0.5	0.3	0.20
Type II discriminant function			
13. Other ground cover	2.1	0.3	0.32**
14. Live stem (<0.5 cm in diameter) density	0.2	0.5	-0.29**
15. Live stem (1.0-2.5 cm in diameter) density	0.1	0.0	0.23*
16. Chamise live leaf density	0.5	0.3	0.31**
17. Vertical canopy density 0-25 cm aboveground	0.2	0.3	-0.43**
18. Live leaf density	0.7	0.5	-0.25*
19. Total stem (0.5-1.0 cm in diameter) density	0.1	0.0	0.31**

*P < 0.05, **P < 0.01

diameter was less. This combination of variables was associated with 21% of the variability in woodrat occurrence (tables 1 and 2).

The only significant partial correlation between woodrat presence and a habitat variable was for amount of other ground cover. Other ground cover consisted mainly of large downed branches which might have attracted the woodrats. Linsdale and Tevis (1951) stated: "Every pile of dead wood within the rat habitat eventually receives an accumulation of twigs deposited by the rats."

Since no significant partial correlation existed between woodrat presence and any other habitat variable and since we cannot explain why other habitat variables in the Type I discriminant function would be particularly important to woodrats, we hypothesize that other significant differences between where woodrats occurred were the result of woodrat activity in their micro-

habitats rather than because woodrats select microhabitats with those particular characteristics. The lower density of live stems less than 0.5 cm in diameter may have resulted from woodrat utilization of the vegetation. Woodrats cut the terminal branches of chamise and wedgeleaf ceanothus and take them to their houses where they feed on leaves and flowers (Linsdale and Tevis 1951). This continual pruning of the plants may cause improved vigor so that a greater density of live chamise leaves and a greater density of live stems between 1.0 cm and 2.5 cm in diameter occur in microhabitats used by woodrats. In this habitat dominated by mature chamise and wedgeleaf ceanothus plants, it appears, then, that woodrat selection of microhabitats may have been in response to other ground cover, while other significant differences between where woodrats did and did not occur were the result of woodrat activity.

The Type II discriminant function for the 25 year-old plot contained four significant and three nonsignificant univariate variables. Again, we could not explain why woodrats would select microhabitats with such specific densities of live leaves, stems between 0.5 cm and 1.0 cm in diameter, or a vertical canopy density between 0 cm and 25 cm above ground, and therefore concluded that this combination has little ecological validity even though more of the variation in woodrat presence was associated with that combination of variables. The combination of variables probably only had statistical significance.

CONCLUSIONS

Discriminant analysis, multiple regression, and partial correlation results can be used together to characterize the microhabitats used by individuals of a species and to generate hypotheses explaining why animals occur in some microhabitats and not in others. The specific form of the hypothesis will vary depending on the type of discriminant function.

If the habitat variables are highly multicollinear, however, estimates of the regression coefficients are unstable and unreliable. In such cases VIFs can be calculated or ridge regression can be used to identify the unstable variables for elimination prior to running the multiple regression and partial correlations.

The multivariate approaches in wildlife studies can suggest hypotheses to explain observations made by the researcher. These hypotheses should themselves then be tested so that advances can be made in our understanding of the habitat relationships of wildlife species.

ACKNOWLEDGMENTS

We thank L.A. Marascuilo for his assistance in the statistical interpretation of our data and we thank D.R. McCullough, W.E. Waters and D.E. Capen for their helpful comments on an earlier draft of our manuscript.

LITERATURE CITED

Cavallaro, J.I. 1978. Small mammal habitat in different-aged chamise/wedgeleaf ceanothus chaparral. M.S. Thesis. 112 p. University of California, Berkeley.

Chatterjee, S., and B. Price. 1977. Regression analysis by example. 342 p. John Wiley and Sons, New York, N.Y.

Dueser, R.D., and H.H. Shugart, Jr. 1978. Microhabitats in a forest-floor small mammal fauna. Ecology 59:89-98.

Hoerl, A.E., and R.W. Kennard. 1970. Ridge regression: applications to non-orthogonal problems. Technometrics 12:69-82.

Holbrook, S.J. 1978. Habitat relationships and coexistence of four sympatric species of Peromyscus in Northwestern New Mexico. Journal of Mammalogy 59:18-26.

Linsdale, J.M., and L.P. Tevis, Jr. 1951. The dusky-footed woodrat. 664 p. University of California Press, Berkeley, Calif.

M'Closkey, R.T. 1976. Community structure in sympatric rodents. Ecology 57:728-739.

M'Closkey, R.T., and B. Fieldwick. 1975. Ecological separation of sympatric rodents (Peromyscus and Microtus). Journal of Mammalogy 56:119-129.

Williams, W.A., C.O. Qualset, and S. Geng. 1979. Ridge regression for extracting soybean yield factors. Crop Science 19:869-873.

DISCUSSION

LESLIE MARCUS: 1) The three types of discriminant analyses, while simpler, remind me of original suggestions in use of uncorrelated variables in physical anthropology (Pearson coefficient of racial likeness). However the type II variables may represent a "gestalt" or common habitat factor for the organism. One way of getting at that might be a rotation of the multivariate discriminant coefficient akin to oblique factor rotations, of course only reasonable for $K > 2$ groups. 2) I would suggest a better choice for your "types" would be the Mahalanobis distance d_i for each variable rather than a significance test. The "type" test is both sample size dependent and dependent on significance and level and, of course, still requires a cutoff for D. 3) I would prefer Mahalanobis D^2 as a measure of difference rather than R^2 for overall discrimination, because it gives a plottable metric in the canonical variate of discriminant space. 4) It is suggested that correlation between canonical variate and original variate be looked at. It has been shown that for two groups this is just a simple function of the mean difference (Bergmann, R.E. 1970. Interpretation and use of a generalized discriminant function. p. 35-60. In Bose, R.C. et al., editors. Essays in probability and

statistics. University of North Carolina Press.) In K groups, this is not so since vectors are not orthogonal. Perhaps for full interpretation these correlations should be given; they are almost never reported.

JANET CAVALLARO: We find your comments on discriminant analysis valuable and we think that they may be very useful in further interpreting discriminant functions. In response to your fourth comment, in the two-group discriminant case, the correlation between the canonical variate and the original variables provides no new information and cannot be used in interpreting the original variables as you stated. In discriminant analysis where the number of groups exceeds two, the correlation may be useful; but it is useful only in trying to interpret what the canonical variate represents, not in interpreting the importance of the original variables in separating the groups. An original variable may be important because it separates the groups or it may be important because it in combination with the other original variables creates a new variable, the canonical variate, which separates the groups.

KEN MORRISON: In your first example, it was only for Peromyscus maniculatus (the most numerous species) that the discriminant function was successful. Does this really mean that the discriminant function failed?

JANET CAVALLARO: A discriminant function must be evaluated both statistically and ecologically; and although it may be significant statistically, we contend that it may be ecologically meaningless. In the case of the discriminant functions reported by Holbrook (1978), we assume that at least some of the discriminant functions must have been statistically significant, but we question whether or not they were ecologically meaningful. She provided no ecological interpretation of the functions except to infer that "species occupy habitats with characteristic three-dimensional structure". We found it surprising ecologically that the microhabitats used by Peromyscus maniculatus could be classified better than those used by P. truei, P. boylii, P. dificilis.

In our work in the chamise-ceanothus chaparral, P. maniculatus and P. truei occurred on all study plots. We could separate those microhabitats P. maniculatus used from those they did not use on only one of those plots. In contrast, on all the study plots, we could separate the microhabitats P. truei used from those they did not use. Considering the ubiquitous distribution of P. maniculatus, these results conform more with current understanding of the habitat relations of these species. We should mention, however, that as researchers we are searching for new information so that the discriminant functions Holbrook reported should not necessarily be negated. We believe, however, that the burden rests on her to hypothesize why the microhabitat used by P. maniculatus could be classifed more readily than those used by the other three species.

GERALD SVENDSEN: Because many of your measured variables are not independent from one another, do you think that you would have gained any insight by using a factor analysis to derive new composite yet independent variables and then performing multiple regression on these new variables?

JANET CAVALLARO: I think the question you are asking is whether or not principal components could have been used to create a set of independent "habitat" variables for the multiple regression analysis rather than using ridge regression to identify redundant or highly intercorrelated habitat variables. The answer is definitely yes, but the use of principal components can create problems in itself. Primarily, each principal component must be interpreted as to what it measures, and this interpretation itself can be a difficult task. By using the ridge regression to identify a set of relatively independent habitat variables, that problem is avoided.

In our approach to analyzing the data, we used the discriminant analysis to characterize the microhabitats used by a particular species and to determine how much of the variance between where a species did and did not occur could be accounted for by the combination of variables in the discriminant function. We were, therefore, interested in the suite of characters which described the microhabitats used by a species and we were not concerned about whether or not the variables were intercorrelated.

The variables in the discriminant function which characterized the species' microhabitats, however, may or may not be the variables to which the animals respond. The subsequent multiple regression and partial correlation analysis was then used to hypothesize why the animals occurred in particular microhabitats and not in others. For this phase of the analysis, however, independent habitat variables were required. We preferred ridge regression over principal components to obtain a set of independent habitat variables because, as already mentioned, we were able to use the habitat variables which were actually measured rather than using statistical habitat variables, principal components, which themselves require interpretation.

PAUL GEISSLER: If I understand what you said, the difference between the conclusions associated with the types of discriminant functions is that significant univariate variables can be interpreted individually while nonsignificant variables cannot. It seems to me that whether or not variables can be interpreted individually depends on whether or not it is correlated with other variables, and not whether or not a relationship can be demonstrated with the response variable. Also it seems to me that limiting the discriminant functions to significant variables may cause one to miss complex relationships which depend on more than one variable.

JANET CAVALLARO: The different types of

discriminant functions do not determine whether or not the individual variables in the discriminant function can be interpreted; the different types of discriminant functions determine how the individual variables must be interpreted. The discriminant function is a new "habitat" variable which is a sum of weighted habitat variables, not a response variable. If the discriminant function contains only significant univariate variables (Type I), the discriminant function can be interpreted as separating microhabitats which have more of variable X, less of variable Y, and more of variable Z from microhabitats which have less of variable X, more of variable Y, and less of variable Z. If the discriminant function contains insignificant variables as well (Type II), the discriminant function can be interpreted as above plus the fact that in microhabitats used by a particular species, variable A has a certain average value. The hypothesis generated from a Type II discriminant function, therefore, may include the idea that variable A, the insignificant univariate variable, must be at a certain threshold level before the animals of a species will use a certain habitat. If that is the case, however, a significant difference should exist in variable A between where animals do and do not occur if a broader range of habitats were sampled. Thereafter, a Type I discriminant function may well be developed to separate the microhabitats used and not used.

In our study in the chamise-ceanothus chaparral, ceanothus density did not differ between where Peromyscus truei did and did not occur on the 25-year-old plot, and yet ceanothus density was a variable in the Type II discriminant function. P. truei occurred where ceanothus density equalled 1.7 in contrast to 1.5 where P. truei did not occur. It was not clear to us why P. truei should select on the average microhabitats with a ceanothus density of 1.7 in contrast to 1.5 when those two values were essentially no different. On the 8-year-old plot ceanothus density was much less, however, and P.truei occurred where the ceanothus density was significantly greater, 1.2 in contrast to 0.2 where they did not occur. From these results on the 8- and 25-year-old plots, we can comfortably hypothesize that P. truei requires a minimum density of ceanothus before they will occupy the chamise-ceanothus habitat. This is analagous to the observations of McCabe and Blanchard (1950) who found that P. truei would not use pure stands of Baccharis pilularis but that they would use the habitat when shrubs of other species were also present. The Type I discriminant function for the 8-year-old plot enables us to develop the above partial hypothesis much more easily than did the Type II discriminant function for the 25-year-old plot.

By restricting the discriminant function to significant univariate variables, complex relationships between animals and their habitats potentially can be missed; but as we have suggested, this will depend to a large extent on the range of habitats sampled. If a broad enough range of habitats is sampled, any complex relationship should be detected through a Type I discriminant function as well as through a Type II or III discriminant function. In addition, we need to consider whether or not a Type II or III discriminant function will necessarily enable us to better detect complex relationships between habitat variables and a particular species. The Type II or III discriminant function suggests that animals are using very precisely defined habitats and so for those discriminant functions to have ecological meaning, we think the measured habitat variables must be very close measures of variables to which the animals are actually responding. We question how often variables perceived by animals when they make their habitat selection are actually measured; and we therefore question whether or not many of the Type II or III discriminant functions are not just artifacts of our sampling.

A DESCRIPTIVE MODEL OF SNOWSHOE HARE HABITAT

Kathryn A. Converse[2] and Bernard J. Morzuch[3]

Abstract.--The snowshoe hare (Lepus americanus) on the southern edge of its range in Massachusetts provided an opportunity for use of multivariate analysis to model selection of common habitat components over an extensive area during the critical winter months for comparison with preferred habitat. Ten areas were chosen for variability in their vegetation structure and composition within two ecological zones of the Berkshire Mountains in western Massachusetts. Track counts were used as indices of hare distribution and regressed on measurements of ground cover; shrub volume; shrub and sapling frequency, richness and composition; tree composition, frequency and basal area; and snow depth.

A linear regression model was used to analyze these hypothesized relationships. The technique of principal components is suggested as a method of dealing with collinearity among independent variables. Tests were performed to judge the appropriateness of pooling data from different study areas. Seemingly unrelated regression was in turn, used to analyze the possibility of mutual correlation within the error structure.

Key words: Forest habitat; Massachusetts; multicollinearity; pooling cross section data; principal components; regression analysis; seemingly unrelated regression; Snowshoe hare.

INTRODUCTION

Massachusetts has over 5 million acres of land with 3 million acres of forests classfied into 40 types. In 1971-1972, 76% of these forests contained trees greater than 12 m in height and at present 45% of the state is harvestable (MacConnell 1975). In addition to increasing forestry operations, there are continuing losses of wildlife habitat due to clearing for housing, commercial and recreational developments, and drainage and filling of wetlands. It is essential that wildlife managers become concerned with the effects of such practices and develop methods for location and description of extensive areas of wildlife habitat for game and nongame species. One possible method that addresses these issues is division of the state into ecological zones (Sczerzenie 1980).

The snowshoe hare (Lepus americanus) is diversely distributed across Massachusetts from the oak-pitch pine barrens of Cape Cod to the spruce-fir thickets of the Berkshire Mountains. Some hare may have survived transplantation from

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]M.S. Candidate, Wildlife Biology, Department of Forestry and Wildlife Management, University of Massachusetts, Amherst, MA 01003.

[3]Assistant Professor, Department of Food and Resource Economics, University of Massachusetts, Amherst, MA 01003.

New Brunswick, a management plan in practice since 1919. It is assumed that survival of these hare was a function of inherent selective mechanisms for corresponding habitat components. Previous studies have shown hare to use a variety of conifer and hardwood stands of all ages, species, and interspersions. In a marginal southern range such as Massachusetts, it is difficult to specify precisely the variables to be used in a habitat model.

This paper discusses a study conducted May 1978 to March 1980 where the physical features of a variety of forest types within two ecological zones in western Massachusetts were sampled intensively. A linear regression model was used to analyze the hypothesized relationships between hare activity and selection of common habitat components during the critical winter months.

STUDY AREAS

Known hare habitat was determined through correspondence and interviews with district wildlife managers, hare hunters, and local residents. Ten research areas were chosen along a 90-km circuit through eight towns located in Berkshire, Hampshire, and Franklin counties in western Massachusetts. Five areas were chosen within each of two ecological zones--transitional (zone 1) and central Berkshires (zone 2)--designated by physiographic, vegetative, and demographic variation (Sczerzenie 1980) (fig. 1). Suitability of all areas was determined by field investigation of size, permission of private land owners, year-round road access, distance from residential or recreational disturbance, vegetation structure and composition variability, presence of browse and/or fecal pellets, and ability to census all areas on snowshoes within 24 hours.

The most recent land use inventory (MacConnell 1975) indicated that the zone 1 study areas fell within predominately hardwood forests, while zone 2 study areas were predominantely softwood forests (table 1). Forest maturity was confirmed in both zones by the 12.5-18.6 m height class and by high density, 81-100% crown closure. When the study areas were viewed separately, forest types of documented importance to the hare, e.g., spruce, became more significant. Discontinuity within habitat types was not usually reflected on land use maps, but the resulting edges provided juxtaposition of food and cover.

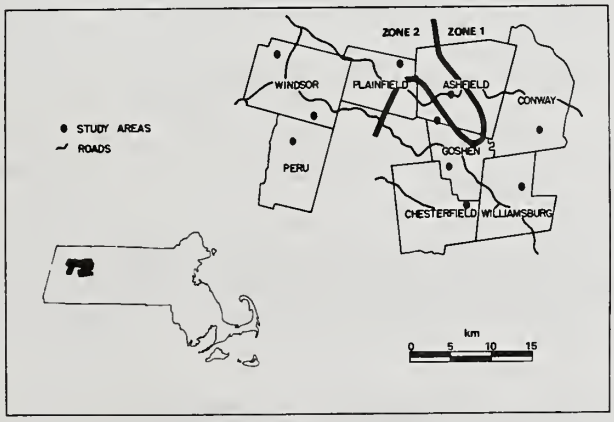


Figure 1. Location of the ten research areas within the eight Massachusetts towns and two ecological zones.

METHODS

Habitat Measurements and Hare Activity

Transect lines (total 400 m) were located randomly on each area with permanent stations every 25 m and plot centers each 50 m. Vegetation was measured September 1978 through July 1979 on all transects at each 50 m plot center. Frequency, composition, and basal area of trees 10 cm dbh and above were measured using a 10 factor prism plot. Circular plots, 40.5 m^2, were used to determine frequency, composition and richness of shrubs 30.5 cm high to 7.5 cm dbh, saplings 7.6 cm to 9.9 cm dbh, and shrub volume for mountain laurel (Kalmia latifolia) and yew (Taxus canadensis). Percent ground cover less than 30.5 cm high was visually estimated on four 2 m x 2 m quadrats and then averaged.

An understory density screen 1 m wide by 2 m high (Telfer 1974, Oldenmeyer 1975) gridded into 5% blocks was placed 15 m from and at right angles to the transect line (Nudds 1977). The screen was photographed and the lower (0-1 m) and upper (1-2 m) percent obscured by vegetation recorded as maximum density in July and minimum in March. Habitat variables used in the analyses are presented below:

RICH = number of species in the shrub and sapling plot,
EVER = evergreen trees >10 cm dbh/0.01 ha,
HARD = hardwood trees >10 cm dbh/0.01 ha,
BA = basal area m^2/ha,
SHRUB = shrub volume 100 m^3/ha,
EDBHO = evergreen stems >30.5 cm high, up to 2.5 cm dbh/0.01 ha,
HDBHO = hardwood stems >30.5 cm high, up to 2.5 cm dbh/0.01 ha,
EDBH1 = evergreen saplings 2.6-9.9 cm dbh/0.01 ha,
HDBH1 = hardwood saplings 2.6-9.9 cm dbh/0.01 ha,
MINL = lower percent minimum understory,
MINU = upper percent minimum understory,
MAXL = lower percent maximum understory,
GC1 = percent ground cover in annuals,

GC2 = percent ground cover in perennials,
GC3 = percent bare ground (leaves, duff and rock).

Hare tracks intersecting the transect line were totalled and snow depth measured for each 50 m section after each fresh snowfall (Hartmann 1960, Brocke 1975). Problems of high winds drifting fresh snow, and the change from snow to sleet along the altitudinal gradient eliminated many observation days. If any sections of a transect line were not trackable the count for the whole line was not used.

Track counts were used as indices of hare cover preference because they can be measured easily by one person, provided for analysis by habitat, and are good for extensive studies where only relative population levels are needed. Tracks provide an index to the average number of hare using an area part of the time or as part of their home range (Adams 1959, Hartman 1960).

Statistical Model

Ordinary Least Squares

Multivariate ordinary least squares (OLS) regression techniques were used to determine the relationships between hare activity as a dependent variable and weather and habitat measures as independent variables.

The hypothesized relationship was assumed to be linear and of the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots\ldots + \beta_k X_{ik} + \epsilon_i \qquad (1)$$

where Y_i = the i^{th} value of the entire set of n observations on the dependent variable Y;

X_{ij} = the i^{th} value of the j^{th} independent variable X_j, where j ranges from 1 to k;

Table 1. Dominant species of forest layers and corresponding forest classification system (MacConnell 1975) for zones 1 and 2.

Research areas	Forest[1] types	Tree[2] >10 cm dbh	Saplings 7.5-10 cm dbh	Shrubs up to 7.5 cm dbh
		Zone 1		
0	HS3A SH3A	Bi H Be	Wh Be M	L Wh Be
3	S3A SH3A	H M	S H	H S
4	H3A SH2A	O H	Wh M Be	L Be Wh
8	SH2A HS3A	H M Wp	M	Bl M L H
9	HS3A	M H Bi	Be H	L Wo H M
		Zone 2		
1	SH2A HS3A SH3A	H M S	M H	L Y M
2	SH2A SH3A	Wp S M	M Wp	Sb M S
5	S3A SH3A HS3A	S	M S	M S Bl
6	HS3A HS2A H2A	M Bi H	Bi S M	M S
7	S3A SH2A SH3A	S	S F	S Bl M F

[1] Type: H = hardwood and S = softwood, 80% stands; mixed HS = hardwood and SH = softwood predominating
Height: 2= 6.4-12.4 m; 3= 12.5-18.6 m
Crown closure: A= 81-100%; B= 30-80%

[2] H=hemlock, M=maple, F=fir, S=spruce, Bi=birch, O=oak, Be=beech, Wp=white pine, Wh=witch hazel, L=laurel, Bl=blueberry, Wo=witch hobble, Sb=steeplebush, Y=yew.

β_j = the j^{th} unknown parameter to be estimated;

β_o = a constant term;

ε_i = the i^{th} value of the set of n random disturbances about the mean of Y.

Equation (1) can be conveniently expressed in matrix notation as

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad (2)$$

where $\underline{Y}$ = an n x 1 observation vector;

X = an n x k matrix of independent variables;

$\underline{\beta}$ = a k x 1 parameter vector on X;

$\underline{\varepsilon}$ = an n x 1 vector of error terms.

Assumptions of the model are that X is of rank k and that each ε_i is normally and identically distributed, independent of the other disturbance terms, and has mean zero and constant variance σ^2.

Violation of any of these assumptions will render OLS estimation inadequate. Specifically, difficulty with respect to the rank condition of the X matrix, i.e., multicollinearity, requires the implementation of some "treatment." Likewise problems with respect to the error term, e.g., heteroscedasticity or autocorrelation, require some manipulative corrective measure to the data before OLS can be applied.

Pooling Data

A very real problem when dealing with cross-section data and model specification in this case is the matter of organizing the data between the two zones for estimation purposes. That is, the model for each zone relating hare activity to weather and habitat measures can be expressed generally as in equation (2), and the question revolves around the manner of implementing the entire set of data from both zones to get the most efficient parameter estimates.

Let equation (2) be rewritten to represent zone 1 as

$$\underline{Y}_1 = X_1\underline{\beta}_1 + \underline{\varepsilon}_1 \qquad (3)$$

where equation (3) is identical to equation (2) except that the subscript 1 relates to zone 1 information. Correspondingly, the model representing zone 2 is

$$\underline{Y}_2 = X_2\underline{\beta}_2 + \underline{\varepsilon}_2. \qquad (4)$$

Estimating the model for zone 1 and zone 2 individually by OLS amounts to the following formulation:

$$\begin{bmatrix} \underline{Y}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & X_2 \end{bmatrix} \begin{bmatrix} \underline{\beta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\beta}_2 \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\varepsilon}_2 \end{bmatrix} \qquad (5)$$

A most important feature of this model is that arranging the data in this fashion does not restrict the parameters of the independent variables in zone 1 to be equal to those in zone 2, i.e., there are 2k parameters to be estimated in the unrestricted model represented by equation (5).

The inclination of many researchers in such a situation is to "stack" equations (3) and (4) as:

$$\begin{bmatrix} \underline{Y}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_2 \end{bmatrix} \underline{\beta} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\varepsilon}_2 \end{bmatrix} \qquad (6)$$

Notice that estimation of this model imposes severe resrictions upon the two parameter vectors of equations (3) and (4), i.e., $\underline{\beta}_1 = \underline{\beta}_2 = \underline{\beta}$. Specifically, the implication is that any particular parameter β_j in zone 1 is equal to the corresponding β_j in zone 2. It is suggested that, rather than estimating only the stacked model of equation (6) and naively imposing severe restrictions on the parameter space, alternative model specifications ought to be tested against each other and one judged more appropriate than another on the basis of some statistical criterion. One such criterion is the classical F-test (Fisher 1970).

Briefly, the idea is to calculate the residual sum of squares of the unrestricted model (RSS_U) represented by equation (5) and make comparisons with the residual sum of squares of the restricted model (RSS_R) in equation (6). The format for making these comparisons is:

$$F = \frac{(RSS_R - RSS_U) \,/\, (DF_R - DF_U)}{RSS_U \,/\, DF_U}, \qquad (7)$$

which is distributed as central F (centrality parameter equal to zero) with $(DF_R - DF_U)$ and DF_U degrees of freedom for the numerator and denominator, respectively. Notice, furthermore, that DF_R and DF_U are the degrees of freedom

associated with the restricted and unrestricted models, respectively.

Under the null hypothesis that $\underline{\beta}_1=\underline{\beta}_2=\underline{\beta}$, i.e., that the restrictions suggested by equation (6) hold, versus the alternate hypothesis that $\underline{\beta}_1 \neq \underline{\beta}_2$, i.e., that the restrictions do not hold, equation (7) has particular appeal. Simply stated, if RSS_R and RSS_U are "close" to each other, the implication is that the restrictions are not in conflict with the sample data. This is reflected in a small calculated F-statistic relative to a critical F such that the null hypothesis cannot be rejected. If, on the other hand, RSS_R diverges greatly from RSS_U, the calculated F-statistic will exceed the critical F and the null hypothesis will be rejected, implying that the restrictions are in conflict with the sample data.

Seemingly Unrelated Regression

Once either the stacked or unstacked model has been selected as the appropriate specification, an important issue revolves around whether or not the OLS procedure used for the estimation of the selected model leads to the most efficient parameter estimates. That is, even though 1) the system has been purged of any autocorrelation or heterscedasticity relating to the error structure and 2) possible restrictions relating to the parameters have been tested, additional information may still be gleaned from the estimated error vectors $\underline{e}_1$ and $\underline{e}_2$. In fact, up to now, the assumptions relating to the error terms of either model (5) or model (6) have been very restrictive in the sense that they do not allow for mutual correlation between $\underline{e}_1$ and $\underline{e}_2$. Testing whether or not the error vectors of the selected model (5) or (6) above are actually unrelated or only seemingly unrelated involves a generalized least squares procedure (Zellner 1962). Basically, testing for mutual correlation involves comparing the estimate of the variance of the selected model (5) or (6) above (call this estimate $\hat{\sigma}^2_{OLS}$) with the estimate of the variance of that same model cast in a generalized least squares or seemingly unrelated regression framework (call this estimate $\hat{\sigma}^2_{SUR}$). The hypothesis to be tested is one of homogeneity of variance with the null hypotheis being $\hat{\sigma}^2_{OLS} = \hat{\sigma}^2_{SUR}$ and the alternate hypotheis being $\hat{\sigma}^2_{OLS} \geq \hat{\sigma}^2_{SUR}$. The appropriate test statistic is

$$F = \hat{\sigma}^2_{OLS}/\hat{\sigma}^2_{SUR}$$

If $\hat{\sigma}^2_{OLS}$ is much larger than $\hat{\sigma}^2_{SUR}$, the null hypothesis of no mutual correlation must be rejected.

The appeal of the seemingly unrelated regression approach is that the off-diagonal elements of the variance-covariance matrix of the error term are not restricted to zero (as in the ordinary least squares case) due to possible mutual correlation. Through the generalized least squares algorithm, the degree of mutual correlation in the error structure is analyzed. If there exists little or no mutual correlation, the zero restrictions on the off-diagonals hold, and $\hat{\sigma}^2_{SUR}$ approaches $\hat{\sigma}^2_{OLS}$. On the other hand, the existence of mutual correlation in the error vectors implies the admissibility of additional information about the system which in turn implies that $\hat{\sigma}^2_{SUR} < \hat{\sigma}^2_{OLS}$ and allows for more efficient parameter estimates.

Principal Components

Collinearity among the independent variables can be detected by observing rather high pairwise correlations between explanatory variables and also by performing auxiliary regressions of individual independent variables on the remaining set. This latter test is suggested by Farrar and Glauber (1967).

The way to deal with multicollinearity is to introduce additional sample information to hopefully increase the selective variation among the independent variables. Such information is normally added by way of restrictions on the parameters suggested by theory. These restrictions may take the form of exact linear restrictions (Goldberger 1964: 256-8), inequality restrictions (Judge and Takayama 1966), and stochastic restrictions (Theil and Goldberger 1961).

In the absence of any theoretical basis for admitting restrictions on the parameters, an alternative is to place restrictions on the independent variables themselves. This can be accomplished by transforming the original variables into artificial constructs that are orthogonal to each other and then retaining certain of these constructs in a regression model on the basis of their contribution to variability in the original data, while eliminating--placing zero restrictions on--those constructs that contribute little or nothing to the variability in the original data. This is precisely the focus of principal components.

In principal components regression, a transformed variable is determined to be important

and included or unimportant and excluded in the regression model depending upon the size of the characteristic root (eigenvalue) associated with its corresponding characteristic vector (eigenvector) (Massy 1965), the statistical significance of its regression coefficient (Mittelhammer and Baritelle 1977), or some combination of eigenvalue size and correlation with the dependent variable (Johnson et al. 1973).

A complication results with respect to the optimal number of components to delete. The tendency traditionally has been to delete components asociated with small eigenvalues, e.g., less than one. The limitation of this approach is that components with small eigenvalues may be correlated very highly with the dependent variable. Thus, a structural norm which simultaneously considers the amount of variability accounted for by a particular component and its correlation with the dependent variable has greater appeal. A particular norm which accounts for these two measures is the F-test described in equation (7). Components can be sequentially deleted until a new restriction, i.e., the deletion of an additional component, causes no improvement with respect to the fit of the equation. Finally, once the "optimal" number of components has been deleted, the principal component estimators can easily be transformed to estimators on the original independent variables (Johnson et al. 1973).

RESULTS AND DISCUSSION

OLS regression models were estimated individually for zone 1 (n=384) and zone 2 (n=480) in the spirit of equation (5) and with the zones stacked (n=864) as suggested by equation (6). Preliminary to performing these regressions, any problems with respect to heteroscedasticity among the error variances of the individual areas were corrected. The model specifications were tested using equation (7) to determine if the parameter restrictions held. The F-values were significant using the central F-test. This separation was consistent with development of these ecological zones and indicates that there were major habitat differences.

Seemingly unrelated regression (SUR) was done on the unstacked data and the possibility of mutual correlation explored. The disturbances in the two equations were found to be not mutually correlated, and the SUR parameters estimators were the same as the OLS estimators. The above tests showed that both the set of explanatory variables and disturbances between zones were not correlated, and there was no justification for combining the data.

Multicollinearity among the explanatory variables was indicated by significant relationships in the simple correlations and auxilliary regressions. Almost perfect collinearity existed between each of the three ground cover measurements (GC1, GC2, AND GC3) and the remaining variables. This was expected, as ground cover was a function of the amount of light reaching the forest floor which, in turn, was determined by canopy closure and understory. Basal area (BA), a measure of canopy closure, was also collinear with the rest of the variables. Furthermore, collinearity among the independent variables was confirmed in that the OLS regressions for each zone had high standard errors resulting in non-significant t-statistics.

Principal components regression (PCR) was done on each zone. Equation (7) was used as the deletion criterium with one component deleted from zone 1 and four from zone 2. Because information was removed from the models, the t-tests were recognized only as indicators of more precise estimators and suggest variables that may be most important habitat descriptors (Freund 1974). Final PCR model coefficients and significance tests appear in tables 2 and 3. The fact that zone 1 was an area of transition between the Connecticut River Valley lowlands and upland Berkshire Mountains indicated that it would be difficult to find common parameters within this zone's heterogeneity.

The final model for zone 1 had eleven significant effects while zone 2 had eight. Three variates of major importance were retained with the same sign in both models. Species richness (RICH), indicating open areas, and basal area (BA), or canopy closure, had negative effects on hare abundance while hardwood shrub frequency (HDBHO) had a positive effect. Four additional variable effects were significant in both zones but carried opposite signs: number of evergreen (EVER) and hardwood (HARD) trees, number of evergreen saplings (EDBH1), and upper percent winter understory (MINU). Sample data were carefully examined for the possibility that the signs changed in response to a particular sensitivity in habitat measurements. The number of hardwood trees appears to have a sample frequency that could explain the change in signs (fig. 2), but because zone 1 already had more hardwood trees, the only interpretation was that increasing numbers of hardwood trees in zone 1 is related to increased hare activity, while in zone 2 the inverse held. The histogram in figure 3 demonstrates the relative frequencies for evergreen trees was similar in both zones, as they were for EDBH1 and MINU, so the same effects were expected but not found.

One hypothesis for the different effects between zones is that the low hare track count of 298 for zone 1, compared with 2226 tracks for zone 2, had low information content for selection of variable effects and the quality of the parameter estimates were questionable. When EVER and HARD were plotted on a three dimensional graph (fig. 4) the same frequency distributions were seen but in addition the low track count values of zone 1 were evident.

There are possible explanations for the different variable effects between zones based on

Table 2. Principal components regression model of the relationship between hare activity and selection of habitat in zone 1.

Variable[1] Names	Estimated coefficients	T- ratio[2]
RICH	-0.115	-1.933*
SNOW	0.010	2.292**
EVER	0.340	7.238**
HARD	0.088	2.347**
BA	-0.064	-4.796**
SHRUB	-0.001	-0.321
EDBH0	-0.009	-1.745*
EDBH1	-0.028	-2.471**
HDBH0	0.003	3.189**
HDBH1	-0.002	-0.419
MINL	0.014	3.056**
MINU	-0.016	-2.382**
MAXL	-0.012	-2.472**
GC1	0.001	0.101
GC2	-0.006	-0.679
GC3	-0.001	-0.103
Intercept	1.459	

[1] see text for variable description
[2] ** P<0.01, * P<0.05

Table 3. Principal components regression model of the relationship between hare activity and habitat component selection in zone 2.

Variable[1] Names	Estimated coefficients	T- ratio[2]
RICH	-0.148	-2.337**
SNOW	-0.013	-1.523
EVER	-0.120	-2.240**
HARD	-0.315	-3.732**
BA	-0.040	-2.816**
SHRUB	0.008	0.512
EDBH0	-0.006	-0.968
EDBH1	0.066	3.166**
HDBH0	0.010	2.761**
HDBH1	0.096	5.430**
MINL	-0.012	-1.074
MINU	0.051	2.747**
MAXL	0.007	0.690
GC1	-0.019	-1.294
GC2	-0.020	-1.233
GC3	0.016	1.698*
Intercept	3.181	

[1] see text for variable description
[2] ** P<0.01, * P<0.05

field observation. For example, median snow depths were the same for both zones while the maximum depth was 41 cm deeper on zone 2. Heavier snows in zone 2 had a greater immediate effect on the habitat. Accumulation of snow on tree branches and crowns bent them down within reach for browsing, and often broke them off entirely providing tips, and buds for food on top of the snow. Shrubs and saplings became completely covered with snow in some storms and hare burrowed under them and used these areas for food and shelter. Snowfalls in zone 1 were lighter and melted faster with less accumulation, due to more mature trees and denser canopy intercepting the snow, preventing most tree damage. This relationship between snow and vegetation in the two zones explained the hares' selection of different components in the sapling (EDBH1 and HDBH1) and density (MINL, MINU and MAXL) classes.

Additional hypotheses could be suggested pertaining to differences between the models for each zone, e.g., response to different species of vegetation, interspecific competition between hare and white-tailed deer (Odocoileus virginianus), and the predator-food complex, but the important point has already surfaced. If the parameter estimates of these models had been restricted by pooling the two data sets, none of these differences between zones and ultimately hare habitat would have been qualified. Opposite effects for the same variable between zones would have eliminated detection of some significant habitat descriptors and given poor quality estimates for others. This problem has severe limitations when these data are being used as a basis for designing and implementing management plans. Too often in wildlife research, extensive habitat management plans are based on local intensive studies. Wildlife managers cannot afford to have unrealized negative effects of habitat manipulation surface, by either short-term

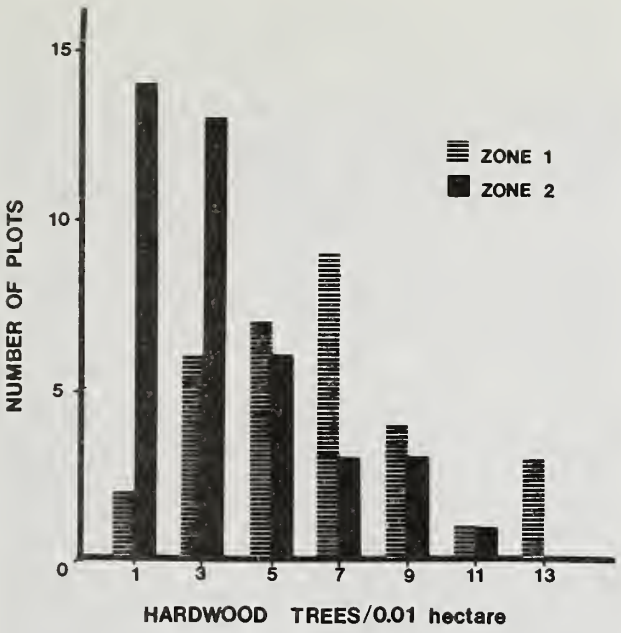


Figure 2. Comparison of frequency distributions of hardwood trees > 10 cm dbh for both zones.

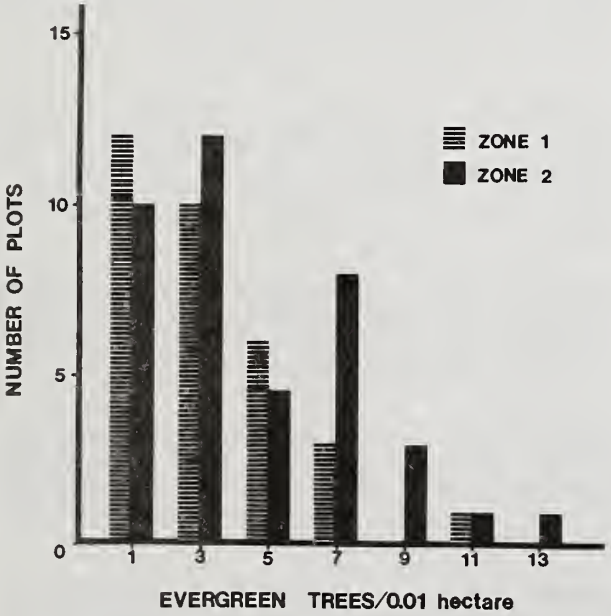


Figure 3. Comparison of frequency distributions between evergreen trees > 10 cm dbh for both zones.

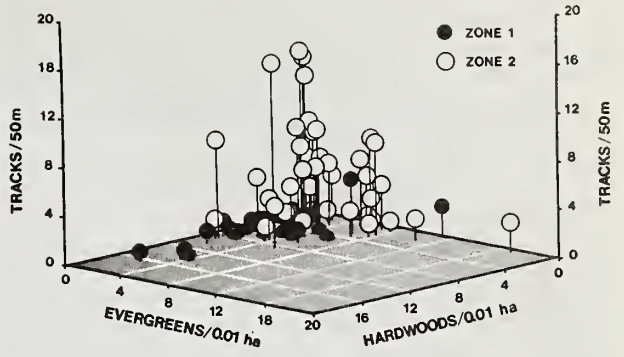


Figure 4. Three dimensional plot of the relationship between hare track counts and number of evergreen and hardwood trees > 10 cm dbh for both zones.

losses of the wildlife in that habitat or long term destruction that will discourage potential immigrating populations.

More specifically, the two zones in this study should be managed differently assuming the parameter estimates are valid. Zone 2 model estimates agreed with our observations of hare activity over the past 2 years, and were consistent with earlier studies in good hare range. The predominately hardwood habitat in zone 1 has only been studied quantitatively recently, and not in areas with the same species composition as this study. Sheldon (1957) stated that northen hardwood-laurel, and hemlock-laurel habitats in zone 1 were the most widespread and important in Massachusetts yet in this study they have the lowest populations. The zone 1 model should be tested with more data to see if the parameter effects hold or will move closer to the other zone model.

ACKNOWLEDGMENTS

This study was jointly supported by the Massachusetts Cooperative Wildlife Research Unit of the U.S. Fish and Wildlife Service, Massachusetts Division of Fisheries and Wildlife, Wildlife Management Institute, and the University of Massachusetts.

LITERATURE CITED

Adams, L. 1959. An analysis of a population of snowshoe hares in north western Montana. Ecological Monographs 29:141-170.

Brocke, R.H. 1975. Preliminary guidelines for managing snowshoe hare in the Adirondacks. Transactions Northeast Section of the Wildlife Society 32:46-66.

Farrar, D.E., and R.R. Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. Review of Economics and Statistics 49:92-107.

Fisher, F.M. 1970. Tests of equality between sets of coefficients in two linear regressions: an expository note. Econometrica 38:361-366.

Freund, R.J. 1974. On the misuse of significance tests. American Journal of Agricultural Statistics 56:192.

Goldberger, A.S. 1964. Econometric Theory. 399 p. John Wiley and Sons, New York, N.Y.

Hartman, F.H. 1960. Census techniques for snowshoe hares. M.S. Thesis. 46 p. Michigan State University, East Lansing, Mich.

Johnson, S.R., S.C. Reimer, and T.P. Rothrock. 1973. Principal components and the problem of multicollinearity. Metroeconomica 25:306-317.

Judge, G.G., and T. Takayama. 1966. Inequality restrictions in regression analysis. Journal of the American Statistical Association 61: 166-181.

MacConnell, W.P. 1975. Classification manual: remote sensing 20 years of change in Massachusetts 1952-1972. 23 p. Massachusetts Agricultural Experiment Station Resource Bulletin No. 631.

Massey, W.F. 1965. Principal components regression in exploratory statistical research. American Statistical Association Journal 60:234-256.

Mittelhammer, R.C., and J.L. Baritelle. 1977. On two strategies for choosing principal components in regression analysis. American Journal of Agricultural Economics 59:336-343.

Nudds, T.D. 1977. Quantifying the vegetative structure of wildlife cover. Wildlife Society Bulletin 5:113-117.

Oldemeyer, J.L. 1975. Characteristics of paper birch saplings browse by moose and snowshoe hares. p. 53-60. In Eleventh North American Moose Conference and Workshop. Winnipeg, Manitoba.

Sczerzenie, P.J. 1980. Ecological zoning of Massachusetts for wildlife management. Transactions of the Northeast Section of The Wildlife Society 37:104-112.

Sheldon, W. 1957. Snowshoe hare in Massachusetts. p. 12-15. In Massachusetts Wildlife, January-February.

Telfer, E.S. 1974. Vertical distribution of cervid and snowshoe hare browsing. Journal of Wildlife Management 38:944-946.

Theil, H., and A.S. Goldberger. 1961. On pure and mixed statistical estimation in economics. International Economic Review 2:65-78.

Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. American Journal of the Statistical Association 57: 350-367.

DISCUSSION

GEORGE BURGOYNE, JR.: Would you care to comment on the justification for treating a large number of observations collected at five sites in two regions as completely independent observations. The fact that only ten sites were used would be expected to yield more homogenous results than if the same number of transects were really located at random. It is difficult to believe that two 50 meter transects adjacently located are as independent as two 50 meter transects independently, randomly located.

KATHRYN CONVERSE: Vegetation frequency, composition, and structure was heterogenous along the ten transect lines. Any problems of the dependence with repeated observations was corrected by orthogonalizing variables using Principle Components. Certainly adjacent 50 m transects are less independent than random 50 m transects, but they did allow for repeated detailed observations of changing hare activity and habitat use with varying weather conditions as the winter progressed. Intensive vegetation measurements and winter tracking over 960 randomly located transects would have been logistically impossible for one person.

KEN MORRISON: Being familiar with animal tracking, I have a gut feeling that the error in measurement increases with increasing activity levels, suggesting that a transformation, perhaps log, would be appropriate? Is it possible that this could have been the case in your study?

KATHRYN CONVERSE: Indeed, a log transformation may be appropriate. We failed to test among competing functional specifications for our model. We merely assumed linear relationships.

KEN MORRISON: The number of crossings will be a function of time since the last snowfall. When you were sampling, is it possible that you consistently sampled one area before the other, thereby creating a biased difference between the two?

KATHRYN CONVERSE: Two sets of five study areas that were a mixture from both zones were tracked in alternating order after each snowfall. I think this would eliminate any bias between zones.

DOUGLAS INKLEY: Do you think that snow depth may have had an effect on hare mobility, and therefore different regional track counts may have been due to both population differences and the effect of snow depth on mobility?

KATHRYN CONVERSE: Regional hare activity was definitely influenced by snow depth, consistency, and accumulation. Inclusion of these snow conditions as variables in this model helped to explain changing food and cover availability by habitat types throughout the winter months.

MARTIN RAPHAEL: How do you derive the individual parameter estimators, i.e., how do you get unique estimators using principal components?

BERNARD MORZUCH: The process is very mechanical and straightforward (See Morzuch, B.J. 1981. Principal components and the problem of multicollinearity. Journal of the Northeastern Agriculture Economics Council. In press.).

LESLIE MARCUS: The "seemingly unrelated regression" application to the two zone hare data seems inappropriate. Since the individual plots between zones (C_{1i} and C_{2i}) are not "contemporaneous elements" (Zellner, A. 1963. Estimators for seemingly unrelated regression equations: some exact finite sample results. American Statistical Association Journal 57:350-367.), there is no reason to relate them. The value of C_1C_2 calculated to estimate σ_{12} depends on the permutation of the M_1 and M_2 observations, and over permutations $E(C_1C_2) = 0$. Since it might happen that $M_1 \neq M_2$ depending on sampling design etc., this is an unnecessary application of the interesting idea. Note if the sampling sites had been sampled in the same order in the two zones or at similar times of the day, this might be a useful test. The way Ms. Converse explained the data collection it does not seem appropriate.

BERNARD MORZUCH: As background to answering this question, two entirely different zones (within which five areas were identifed, with eight plots per area) were delineated on the basis of vegetation and elevation. There was no attempt to match a particular plot in zone 1 against any one plot in zone 2. Alternatively, performing all possible permutations to find one ordering that provided the highest degree of mutual correlation was not the central theme of this analysis.

Relative to the question, who is to say that ordering the cross sections in zones 1 and 2 in a particular fashion to make certain that we have a matching of contemporaneous elements is the one correct ordering so as to make use of Zellner's framework. Given the heterogeneous nature of plots between zones and the lack of theory to suggest a proper matching, alternative orderings, i.e., random or otherwise, are appropriate.

The purpose of the study was to establish a functional relationship between hare distribution and habitat variables. The resulting estimators of the pooled model have, at least, nice asymptotic properties. Therefore, any additional testing of the restrictive assumptions of the OLS model that might lead to greater efficiencies in the estimators would be most welcome. Such is the purpose of testing for mutual correlation in a seemingly unrelated (SUR) framework. Applying SUR in this fashion is, most certainly, appropriate and not an unnecessary application of an interesting idea. Given our ordering of contemporaneous elements, we were unable to conclude the presence of mutual correlation. This is not to say that another ordering would lead to the same conclusions on efficiency grounds. However, consistency and asymptotic unbiasedness of the estimators would remain unaffected under any ordering.

A DISCUSSION OF ROBUST PROCEDURES IN MULTIVARIATE ANALYSIS[1]

Lyman L. McDonald[2]

Abstract.--Robust procedures in multivariate analysis are briefly reviewed. Conjectures are given for applications of existing methods in other areas of multivariate analysis.

Key words: Discriminant analysis; multicollinearity; outliers; principal components; robust procedures.

Upon review of titles of papers to be presented at this meeting, my first reaction was to discuss a combination of three subjects: 1) the danger of "overfitting" multivariate data with descriptive procedures such as principal components, factor analysis and canonical correlation analysis; 2) the ultra-conservative nature of inference procedures in multivariate analysis of variance (MANOVA) and in multivariate regression analysis (i.e., simultaneous regression of several dependent variables on a collection of independent variables); and 3) the need for "robust" analysis procedures in multivariate studies.

The first of my concerns is addressed by Karr and Martin (1981). Their observations reinforce my experiences and I endorse their suggestion that "One possibility might be a table of expected amount of variation accounted for by random number matrices that can be used as a test relative to the amount of variation accounted for in real data." Actually tables of upper percentage points in the null distribution of variance accounted for by random number matrices would be more appropriate for a formal test of hypothesis, but the idea is the same.

The second of my concerns is not of much interest in the papers presented. MANOVA is mentioned rarely and insofar as I am aware, the corresponding simultaneous inference procedures are not strongly recommended.

I emphasize, then, the third topic: a need for robust procedures. This need becomes obvious when reading several of the papers. For example, Williams (1981) correctly pointed out that an important aspect of canonical variates in discriminant analysis is the ecological meaning of the coefficients. However he noted that "Most people who have worked with discriminant analysis have probably seen cases in which positively correlated variates have canonical coefficients with different signs," and quoting from Weiner and Dunn (1966), "For complex structures, magnitudes and even signs of coefficients are dependent on what additional variables are included in the model." In a study of discriminant analysis by use of multiple regression methods, Cavallaro et al. (1981) stated that "Multicollinearity can cause the regression coefficients to be inflated in absolute value or even to have the wrong sign." Smith (1981) in his paper on canonical correlation analysis noted that "Most often coefficients are exceedingly difficult to interpret, if not meaningless (Cassie and Michael 1968). . .." Also, ". . .Cassie (1969) found that adding and subtracting variables at random 'capriciously' changed the values of the corresponding canonical coefficients." Harner and Whitmore (1981) stated that "Outliers in multivariate data can have pronounced effects on the interpretations and conclusions of statistical analyses."

With the exception of Harner and Whitmore (1981) and Cavallaro et al. (1981), little information has been given in these proceedings to help the reader understand and solve these problems when they arise in a particular

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington Vt.

[2]Professor, Departments of Statistics and Zoology, University of Wyoming, Laramie, WY 82071.

application. Instead, the reader is warned that problems may exist and in some cases analysis procedures for detection of the need for more robust methods have been given (e.g., data splitting). However, alternative methods have not always been given if unstable results occur in the analysis of a particular data set.

There appear to be two basic problems and hence at least two approaches to "robust" multivariate analysis procedures: 1) elimination (or reduction) of multicollinearity before inversion of a variance-covariance type matrix--or equivalently, improved estimation of the inverse of a variance-covariance matrix; and 2) robust estimation of parameters by some procedure for outright trimming of "outliers" or for reduction of the influence of outliers.

Consider first the problem of reducing effects of multicollinearity on multivariate procedures. Most multivariate methods, except for descriptive procedures such as principal components and factor analysis, involve the inverse of a matrix, $\underline{S}$, of sums of squares and cross-products. Correlation matrices, variance-covariance matrices, etc., are all special cases. Letting $\lambda_i, i=1,\ldots,p$, denote the characteristic roots of a (p x p) nonsingular matrix $\underline{S}$, one can write

$$\underline{S}^{-1} = \sum_{i=1}^{p} (1/\lambda_i)\, \underline{a}_i \underline{a}_i' \qquad (1)$$

where $\underline{a}_i$ denotes the corresponding normalized characteristic vectors. If some of the variables used to compute $\underline{S}$ are strongly intercorrelated (i.e., there is multicollinearity) then some of the characteristic roots will be "small" and the matrix $\underline{S}$ is said to be ill-conditioned. Clearly from Eq.(1) division by small values is involved and the inverse of $\underline{S}$ will be unstable in such cases. Consider the obvious effect in regression analysis [$\hat{\beta} = \underline{S}^{-1}\underline{X}'\underline{Y}$], discriminant functions [$D(\underline{X}) = (\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'\underline{S}^{-1}\underline{X}$], Mahalanobis' distance [$D^2 = (\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'\underline{S}^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)$] and canonical variables and correlations [characteristic roots and vectors of $\underline{S}_{11}^{-1}\underline{S}_{12}\underline{S}_{22}^{-1}\underline{S}_{21}$ and $\underline{S}_{22}^{-1}\underline{S}_{21}\underline{S}_{11}^{-1}\underline{S}_{12}$] where standard notation is used. The regression coefficients, discriminant coefficients, . . ., will be impossible to interpret if the problem is severe.

Regression analysis has, of course, received the most attention under two general approaches. First, regression on the first few principal components (say q with q < p) eliminating those corresponding to small characteristic roots. The results are obvious in that the reduced coefficient matrix of the normal equations is diagonal and the regression coefficients are

$$\underline{\beta}^* = (\text{diag}\ (\lambda_1,\ldots,\lambda_q))^{-1} \begin{bmatrix} \underline{a}_1' \\ \cdot \\ \cdot \\ \cdot \\ \underline{a}_q' \end{bmatrix} \underline{X}'\underline{Y}\ .$$

Increased stability can be expected with usually only a minor increase in bias. Sczerzenie (1981), in his paper on analysis of deer harvest-land use relationships, correctly noted that the procedure will often reduce the variance of the estimators. Converse and Morzuch (1981) noted that the procedure will decrease the effect of multicollinearity among independent variables. For further details on this procedure, the interested reader is referred to Sczerzenie (1981) and Converse and Morzuch (1981) and their cited references.

Use of principal components to reduce the effect of multicollinearity in multivariate procedures other than regression analysis has been proposed but not studied (insofar as I am aware). The approach is to transform all data to the first few principal components before starting the analysis. Variables are then uncorrelated and only diagonal matrices need to be inverted. For example, Fisher's linear discriminant function

$$D(\underline{X}) = [(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'\underline{S}^{-1}]\underline{X}$$

would be replaced by

$$D^*(\underline{X}) = [(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'\underline{A}')(\underline{A}\underline{S}\underline{A}')^{-1}\underline{A}]\underline{X} \text{ where}$$

$\underline{A}_{qxp}\underline{X}_{px1}$ represents the transformation of all data to the "first" q principal components, q < p. Hopefully, the loadings in $D^*(\underline{X})$ will be more stable and easier to interpret than in the original function without sacrificing the discriminating ability. Similar adjustments can be made in canonical variate analysis and other multivariate procedures.

The second approach to solving the problem of multicollinearity in regression analysis is by ridge regression or biased-regression, first developed by Hoerl and Kennard (1970). Cavallaro et al. (1981) gave an excellent application of this procedure in their use of multiple regression techniques for discriminant analysis. The basic idea is to allow a certain bias in the regression coefficient by adding a constant k to the diagonal of the coefficient matrix of the normal equations before inverting, i.e.,

$$\hat{\underline{\beta}} = (\underline{S}+k\underline{I})^{-1}\underline{X}'\underline{Y}.$$

This operation will add the constant k to all characteristic roots of $\underline{S}$ and hence

$$(\underline{S}+k\underline{I})^{-1} = \sum_{i=1}^{k} [1/(\lambda_i+k)]\underline{a}_i\underline{a}_i' \ .$$

Ridge regression can be viewed as a special application of a much broader problem, namely improved estimation of the inverse of a variance-covariance matirx, e.g., Efron and Morris (1976). In general, small characteristic roots of $\underline{S}$ are underestimates of the corresponding population values and large roots are overestimates. Efron and Morris sought to improve $\underline{S}^{-1}$ by shrinking all roots toward a common value. The ridge-adjusted estimate, $(\underline{S}+k\underline{I})^{-1}$, works by increasing all roots by a constant amount, k. Thus the ill-conditioning effect of underestimating small roots is countered without having much influence on the larger roots. I will refer to estimates of $\underline{\Sigma}^{-1}$ based on adjusted roots of $\underline{S}$ as simply ridge-adjusted estimators.

I am aware of two investigations into the use of ridge-adjusted estimates of $\underline{\Sigma}^{-1}$ in linear discriminant analysis, i.e.,

$$D^*(\underline{X}) = [(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'(\underline{S} + k\underline{I})^{-1}]\underline{X}.$$

The first by DiPillo (1976) showed that when the variance-covariance matrices are ill-conditioned, improvements can be made in the probability of correct classification. In an independent investigation, Smidt and McDonald (1976) found no improvement in classification rate but obtained increased stability in estimates of coefficients of discriminant functions.

I am not aware of research of this type into the study of canonical correlation analysis and other procedures but would conjecture that similar improvements can be made by using ridge-adjusted estimators of the inverse of variance-covariance matrices.

The second area of development of robust multivariate analysis procedures is in the reduction of the effect of outliers on resulting analyses. It is well known that outliers in multidimensional studies are very difficult to detect, and if undetected can have a pronounced effect on the analysis. Harner and Whitmore (1981) developed a method for building robust discriminant models. In general, they found that in the presence of outliers robust models performed much better than Fisher's discriminant model. With such an improvement in discriminant analysis it is almost a certainty that other multivariate procedures will perform better if the influence of outliers is reduced. Guarding against the adverse effects of outliers should be a goal in every data analysis. The interested reader is referred to cited references in Harner and Whitmore's excellent paper for further work in this area.

LITERATURE CITED

Cassie, R.M. 1969. Multivariate analysis in ecology. Proceedings of the New Zealand Ecological Society 16:53-57.

Cassie, R.M., and A.D. Michael. 1968. Fauna and sediments of an interidal mudflat: a multivariate analysis. Journal of Experimental Marine Biology and Ecology 2:1-23.

Cavallaro, J.I., J.W. Menke, and W.A. Williams. 1981. Use of discriminant analysis and other statistical methods in analyzing microhabitat utilization of dusky-footed woodrats. Proceedings of this workshop.

Converse, K.A., and B.J. Morzuch. 1981. A descriptive model of snowshoe hare habitat. Proceedings of this workshop.

DiPillo, P.J. 1976. The application of bias to discriminant analysis. Communications in Statistics A, Theory and Methods 5:834-844.

Efron, B., and C. Morris. 1976. Multivariate empirical Bayes and estimation of covariance matrices. The Annals of Statistics 4:22-32.

Harner, E.J., and R.C. Whitmore. 1981. Robust principal component and discriminant analysis of two grassland bird species habitat. Proceedings of this workshop.

Hoerl, A.E., and R.W. Kennard. 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12:55-67.

Karr, J.R., and T.E. Martin. 1981. Random numbers and principal components: further searches for the unicorn? Proceedings of this workshop.

Sczerzenie, P.J. 1981. Principal components analysis of deer harvest--land use relationships in Massachusetts. Proceedings of this workshop.

Smidt, R.K., and L.L. McDonald. 1976. Ridge discriminant analysis. Technical report No. 108, Department of Statistics, University of Wyoming, Laramie.

Smith, K.G. 1981. Canonical correlation analysis and its use in wildlife habitat studies. Proceedings of this workshop.

Weiner, J.M., and O.J. Dunn. 1966. Elimination of variates in linear discrimination problems. Biometrics 22:268-275.

Williams, B.K. 1981. Discriminant analysis in wildlife research: theory and applications. Proceedings of this workshop.

WORKSHOP PARTICIPANTS

Robert W. Aho
Michigan Department of Natural Resources
8562 East Stoll Road, Route 1
East Lansing, MI 48823

Bruce Ambuel
Department of Wildlife Ecology
University of Wisconsin
Madison, WI 53705

Robert Anthony
Department of Fisheries and Wildlife
Oregon State University
Corvallis, OR 97330

Barbara A. Bajusz
Pesticide Research Laboratory
The Pennsylvania State University
University Park, PA 16802

Pierre Beland
Canadian Wildlife Service
2700 Blv. Laurier
Ste-Foy, Quebec, Canada G1V 4H5

John A. Bissonette
Oklahoma Cooperative Wildlife Research Unit
Oklahoma State University
Stillwater, OK 74074

Jo Black
School of Forest Resources and Conservation
University of Florida
Gainesville, FL 32611

Debbie Boles
Zoology Department
University of Vermont
Burlington, VT 05405

Andre Bourget
Canadian Wildlife Service
2700 Blv. Laurier
Ste-Foy, Quebec, Canada G1V 4H5

Allen Boynton
North Carolina Wildlife Resources Commission
Rt. 5, Box 17
Burnsville, NC 28714

Richard Bramwell
Wildlife Resources
Box 400
MacDonald College
Ste. Anne-de-Bellevue, Quebec, Canada H9X 1C0

Steve Brown
West Virginia Department of Natural Resources
Box 67
Elkins, WV 26241

Francine Buckley
Buckley Associates
372 South Street
Carlisle, MA 01741

Paul A. Buckley
National Park Service
15 State Street
Boston, MA 02129

George E. Burgoyne, Jr.
Michigan Department of Natural Resources
Box 30028
Lansing, MI 48910

M. Gail Butler
Canadian Wildlife Service
Environment Canada
Ottawa, Ontario, Canada K1A 0E7

John Cary
Department of Wildlife Ecology
University of Wisconsin
226 Russell Laboratory
Madison, WI 53706

Gilles Chapdelaine
Canadian Wildlife Service
2700 Blv. Laurier
Ste-Foy, Quebec, Canada G1V 4H5

Carl J. Christianson
University of Maryland
Appalachian Environmental Laboratory
57 First Street
Frostburg, MD 21532

Sukwoo Chang
National Marine Fisheries Service
Highlands, NJ 07732

List does not include authors of papers in these proceedings.

Robert Clark
Wildlife Resources
Box 400
MacDonald College
Ste. Anne-de-Bellevue, Quebec, Canada H9X 1C0

Wayne L. Cornelius
Southeastern Cooperative Fish and Game Statistics Project
North Carolina State University
Box 5457
Raleigh, NC 27650

Brian Dennis
213 Mueller Laboratory
The Pennsylvania State University
University Park, PA 16802

Tim G. Dilworth
Department of Biology
University of New Brunswick
Fredericton, New Brunswick, Canada

Susan M. Doehlert
Department of Wildlife
Division of Forestry
Morgantown, WV 26506

Jean Doucet
Wildlife Resources
Box 400
MacDonald College
Ste. Anne-de-Bellevue, Quebec, Canada, H9X 1C0

Tom Dowhan
Zoology Department
University of Vermont
Burlington, VT 05405

Tom Dwyer
United States Fish and Wildlife Service
Migratory Bird and Habitat Research Laboratory
Laurel, MD 20811

Chuck Evans
USDA Forest Service
2081 East Sierra
Fresno, CA 93710

John T. Finn
Department of Forestry and Wildlife
University of Massachusetts
Amherst, MA 01003

Sidney S. Frissell
School of Forestry
University of Montana
Missoula, MT 59812

J. Edward Gates
Appalachian Environmental Laboratory
Frostburg State College
Frostburg, MD 21532

Jean Gauthier
Canadian Wildlife Service
2700 Blv. Laurier
Ste-Foy, Quebec, Canada G1V 4H5

Paul H. Geissler
U.S. Fish and Wildlife Service
Migratory Bird and Habitat Research Laboratory
Laurel, MD 20811

Arnie Gotfryd
Department of Zoology
University of Toronto
Toronto, Ontario, Canada M5S 1A1

W.E. Grant
Nagle Hall
Texas A&M University
College Station, TX 77843

Bart Guetti
New York State Department of Environmental Conservation
South Wolf Road
Albany, NY 12233

Kevin J. Gutzwiller
The Pennsylvania State University
209B Ferguson Building
University Park, PA 16802

Jonathan Haufler
Department of Fisheries and Wildlife
Michigan State University
East Lansing, MI 48824

Larry Haugh
Statistics Program
University of Vermont
Burlington, VT 05405

Cliff Hawkes
USDA Forest Service
240 West Prospect
Fort Collins, CO 80521

William M. Healy
USDA Forest Service
180 Canfield Street
Morgantown, WV 26505

Doug Heimbuch
Fernow Hall
Cornell University
Ithaca, NY 14853

Jonas Hedberg
Hyltemasa
S-284 00 Perstorp, Sweden

Barbara T. Hill
USDA Forest Service
Box 638
Laconia, NH 03246

Douglas Inkley
U.S. Fish and Wildlife Service
Migratory Bird and Habitat Research Laboratory
Laurel, MD 20811

Rich Kahl
Missouri Cooperative Wildlife Research Unit
University of Missouri
Columbia, MO 65211

Ron Kalinoski
Biology Department
Syracuse University
Syracuse, NY 13210

Marie Kautz
New York State Department of Environmental Conservation
Wildlife Resource Center
Delmar, NY 12054

Jeff Keller
Fernow Hall
Cornell University
Ithaca, NY 14853

C. William Kilpatrick
Zoology Department
University of Vermont
Burlington, VT 05405

M. Kingsley
Canadian Wildlife Service
5320-122 Street
Edmonton, Alberta, Canada

Kevin R. Kinsley
132 Land and Water Resources
The Pennsylvania State University
University Park, PA 16802

J. Thomas Kitchings
Oak Ridge National Laboratory
Building 1505, Box X
Oak Ridge, TN 37830

George LaBar
Wildlife Biology Program
University of Vermont
Burlington, VT 05405

Pierre Laporte
Canadian Wildlife Service
2700 Blv. Laurier
Ste-Foy, Quebec, Canada G1V 4H5

Terry A. Larson
Department of Biological Sciences
Illinois State University
Normal, IL 61761

Richard Lindeborg
USDA Forest Service
240 West Prospect
Fort Collins, CO 80521

David A. MacKenzie
Department of Zoology
University of Toronto
Toronto, Ontario, Canada M5S 1A1

Linda K. Mann
Oak Ridge National Laboratory
Building 1505, Box X
Oak Ridge, TN 37830

Leslie F. Marcus
American Museum of Natural History
79th at Central Park West
New York, NY 10024

Don McCarty
Texas Parks and Wildlife Department
4200 Smith School Road
Austin, TX 78744

Donald A. McCrimmon, Jr.
National Audubon Research Department
Cornell University
159 Sapsucker Woods Road
Ithaca, NY 14850

Eileen Miller
Box 181, RR 2
Dover, NH 03820

Linda Morris
301D Forest Resources Laboratory
The Pennsylvania State University
University Park, PA 16802

Guy Morrison
Canadian Wildlife Service
Ontario Region Headquarters
1725 Woodward Drive
Ottawa, Ontario, Canada K1G 3Z7

Ken Morrison
Department of Chemical Engineering
University of Sherbrooke
Sherbrooke, Quebec, Canada

Ken Myers
Zoology Department
University of Guelph
Guelph, Ontario, Canada

John L. Oldemeyer
Denver Wildlife Research Center
1300 Blue Spruce Drive
Fort Collins, CO 80524

Richard J. Olson
Oak Ridge National Laboratory
Building 1505, Box X
Oak Ridge, TN 37830

John W. Ozard
New York State Department of Environmenal Conservation
Wildlife Resources Center
Delmar, NY 12054

Stephen G. Parren
Wildlife Biology Program
University of Vermont
Burlington, VT 05405

Barbara W. Patty
National Audubon Society
115 Indian Mound Trail
Tavernier, FL 33070

John Paul
Department of Entomology
University of Delaware
Newark, DE 19711

John H. Porter
Department of Environmental Sciences
Clark Hall
University of Virginia
Charlottesville, VA 22903

Dale Rabe
Department of Fish and Wildlife
Michigan State University
East Lansing, MI 48823

Richard L. Raesly
University of Maryland
Appalachian Environmental Laboratory
11 Bowery Street
Frostburg, MD 21532

Gerald Rasmussen
New York State Department of Environmental Conservation
Wildlife Resources Center
Delmar, NY 12054

William Roberts
Wildlife Biology Program
University of Vermont
Burlington, VT 05405

Ken Ross
Canadian Wildlife Service
Ontario Region Headquarters
1725 Woodward Drive
Ottawa, Ontario, Canada K1G 3Z7

Douglas E. Runde
Wildlife Biology Program
University of Vermont
Burlington, VT 05405

Hans Schreuder
USDA Forest Service
240 West Prospect
Fort Collins, CO 80521

Mark Scott
Wildlife Biology Program
University of Vermont
Burlington, VT 05405

Steven L. Sheriff
Missouri Department of Conservation
1110 College
Columbia, MO 65201

Jeffrey Short
United States Air Force
AFESC/DEVN
Tyndall Air Force Base, FL 32403

Nova Silvy
Wildlife and Fisheries Sciences Department
Texas A&M University
College Station, TX 77843

James E. Skaley
Fernow Hall
Cornell University
Ithaca, NY 14853

Charles R. Smith
Laboratory of Ornithology
Cornell University
159 Sapsucker Woods Road
Ithaca, NY 14850

C.E. John Smith
Biometrics Division
Canadian Wildlife Service
Department of the Environment
Ottawa, Ontario, Canada K1A 0E7

Graham Smith
University of Idaho
2712 Pleasanton
Boise, ID 83702

Ronald Smith
Arizona Game and Fish Department
2222 W. Greenway Road
Phoenix, AZ 85023

Alan J. Steiner
Department of Forestry and Wildlife Management
204 Holdsworth Hall
University of Massachusetts
Amherst, MA 01003

Craig Stihler
West Virginia Department of Natural Resources
Box 67
Elkins, WV 26241

Gerald L. Storm
Pennsylvania Cooperative Wildlife Research Unit
The Pennsylvania State University
University Park, PA 16802

Scott Sutcliffe
Loon Preservation Committee
Audubon Society of New Hampshire
RFD 2
Meredith, NH 03253

Gerald E. Svendsen
Zoology Department
Ohio University
Athens, OH 45701

Diane L. Tessaglia
Wildlife Biology Program
University of Vermont
Burlington, VT 05405

Frank R. Thompson
Wildlife Biology Program
University of Vermont
Burlington, VT 05405

Ian Thompson
Canadian Wildlife Service
1725 Woodward Drive
Ottawa, Ontario, Canada K1G 3Z7

Nancy Tilghman
USDA Forest Service
Hilton House
University of Massachusetts
Amherst, MA 01003

Alan Tipton
Department of Fisheries and Wildlife Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24060

Kimberly Titus
University of Maryland
Appalachian Environmental Laboratory
Frostburg, MD 21532

James Traynor
New York State Department of Environmental Conservation
Wildlife Resources Center
Delmar, NY 12054

Walter M. Tzilkowski
205 Forest Resources Laboratory
The Pennsylvania State University
University Park, PA 16802

Beatrice VanHorne
Biology Department
University of New Mexico
Albuquerque, NM 87131

James S. Wakeley
205 Forest Resources Laboratory
The Pennsylvania State University
University Park, PA 16802

Dan Welsh
Canadian Wildlife Service
1725 Woodward Drive
Ottawa, Ontario, Canada K1G 3Z7

Glenn White
6653 Tiffin Avenue
San Diego, CA 92114

Jim Woehr
State University of New York at Plattsburgh
Plattsburgh, NY 12901

☆ U. S. GOVERNMENT PRINTING OFFICE 1981 - 781-065/466 Reg. 8

Capen, David E., editor. 1981. The use of multivariate statistics in studies of wildlife habitat. USDA Forest Service General Technical Report RM-87, 249 p. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo.

This report contains edited and reviewed versions of papers presented at a workshop held at the University of Vermont in April 1980. Topics include sampling avian habitats, multivariate methods, applications, examples, and new approaches to analysis and interpretation.

Keywords: Wildlife habitat, multivariate statistics

Capen, David E., editor. 1981. The use of multivariate statistics in studies of wildlife habitat. USDA Forest Service General Technical Report RM-87, 249 p. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo.

This report contains edited and reviewed versions of papers presented at a workshop held at the University of Vermont in April 1980. Topics include sampling avian habitats, multivariate methods, applications, examples, and new approaches to analysis and interpretation.

Keywords: Wildlife habitat, multivariate statistics

Capen, David E., editor. 1981. The use of multivariate statistics in studies of wildlife habitat. USDA Forest Service General Technical Report RM-87, 249 p. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo.

This report contains edited and reviewed versions of papers presented at a workshop held at the University of Vermont in April 1980. Topics include sampling avian habitats, multivariate methods, applications, examples, and new approaches to analysis and interpretation.

Keywords: Wildlife habitat, multivariate statistics

Capen, David E., editor. 1981. The use of multivariate statistics in studies of wildlife habitat. USDA Forest Service General Technical Report RM-87, 249 p. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo.

This report contains edited and reviewed versions of papers presented at a workshop held at the University of Vermont in April 1980. Topics include sampling avian habitats, multivariate methods, applications, examples, and new approaches to analysis and interpretation.

Keywords: Wildlife habitat, multivariate statistics

Rocky Mountains

Southwest

Great Plains

U.S. Department of Agriculture
Forest Service

Rocky Mountain Forest and Range Experiment Station

The Rocky Mountain Station is one of eight regional experiment stations, plus the Forest Products Laboratory and the Washington Office Staff, that make up the Forest Service research organization.

RESEARCH FOCUS

Research programs at the Rocky Mountain Station are coordinated with area universities and with other institutions. Many studies are conducted on a cooperative basis to accelerate solutions to problems involving range, water, wildlife and fish habitat, human and community development, timber, recreation, protection, and multiresource evaluation.

RESEARCH LOCATIONS

Research Work Units of the Rocky Mountain Station are operated in cooperation with universities in the following cities:

Albuquerque, New Mexico
Bottineau, North Dakota
Flagstaff, Arizona
Fort Collins, Colorado*
Laramie, Wyoming
Lincoln, Nebraska
Lubbock, Texas
Rapid City, South Dakota
Tempe, Arizona

*Station Headquarters: 240 W. Prospect St., Fort Collins, CO 80526